P&S Modern SSDs

Address Mapping & Garbage Collection

Dr. Mohammad Sadrosadati
Prof. Onur Mutlu
ETH Zürich
Fall 2022
23 November 2022

Recap: SSD & NAND Flash Memory

- SSD organization
  - SSD controller: Multicore CPU + per-channel flash controllers
  - DRAM: Metadata store, 0.1% of SSD capacity
  - NAND flash chips
    - Channel (Package(s)) – Die – Plane – Block – Page

- NAND flash characteristics
  - Erase-before-write, asymmetry in operation units (read/program: page, erase: block), limited endurance, retention loss…

- Basic NAND flash opeartions
  - Read/program/erase

Recap: Advanced Commands

- **Subpage Sensing & Random Data Out (RDO)**
  - For I/O-unit mismatch b/w OS and NAND flash memory
- **Cache Read Command**
  - For improving a chip's read throughput
  - By overlapping data transfer and page sensing
- **Multi-Plane Operations**
  - For improving a chip's throughput
  - By enabling concurrently operation of multiple planes
- **Program & Erase Suspensions**
  - For improving the read latency (operation latency asymmetry)
  - By prioritizing latency-sensitive reads over writes/erases

Today's Agenda

- Advanced NAND Flash Commands
- Address Translation & Garbage Collection

Flash Translation Layer: Overview

- SSD firmware (often referred to as SSD controller)
  - Provides backward compatibility with traditional HDDs
  - By hiding unique characteristics of NAND flash memory

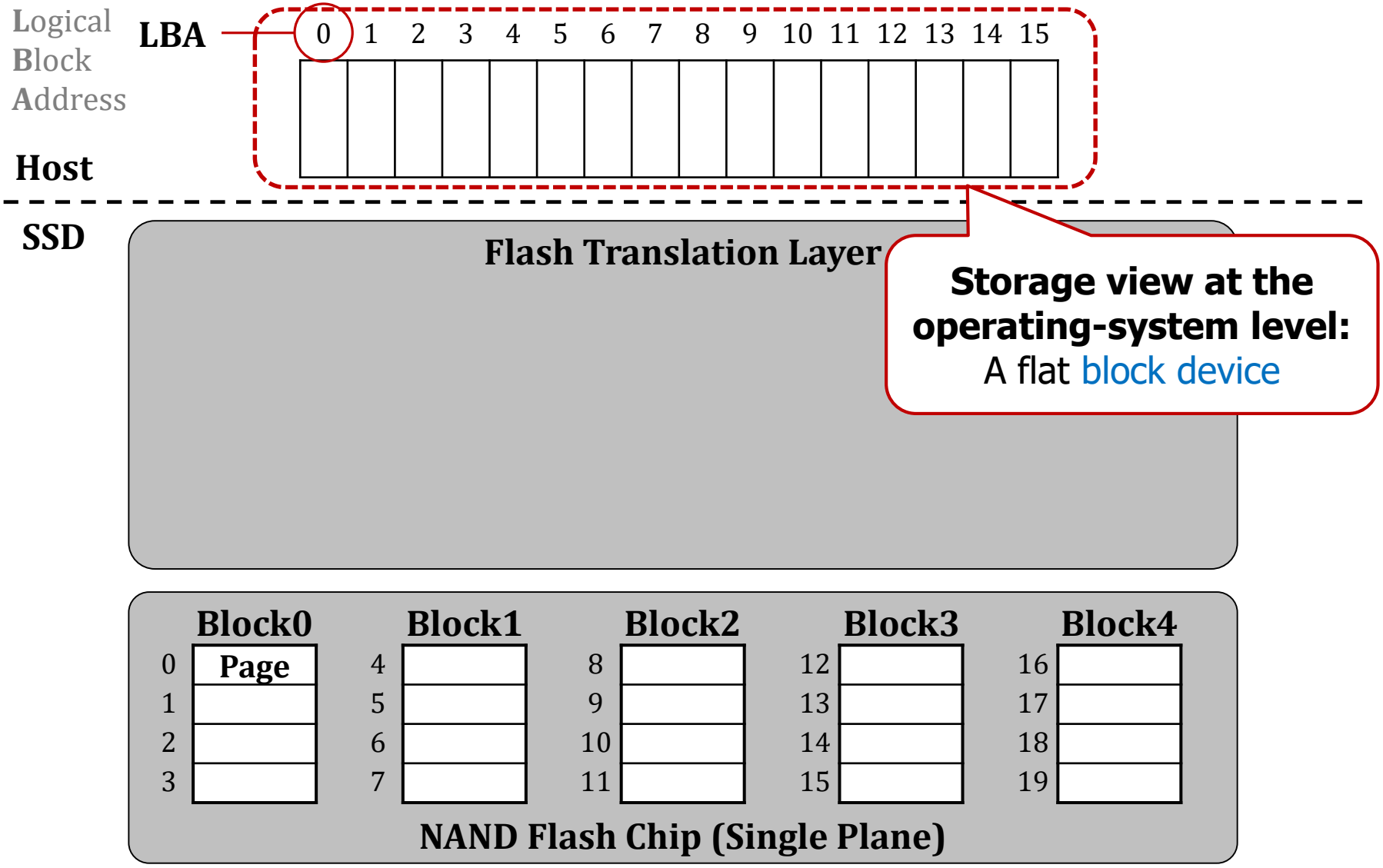
- Responsible for many important SSD-management tasks
  - Address translation + garbage collection
    - Performs out-of-place writes due to erase-before-write property
  - Wear leveling
    - To prolong SSD lifetime by evenly distributing P/E cycles
  - Data refresh
    - Resets transient errors by copying data to a new page(s)
  - I/O scheduling
    - To take full advantage of SSD internal parallelism

Flash Translation Layer: Overview

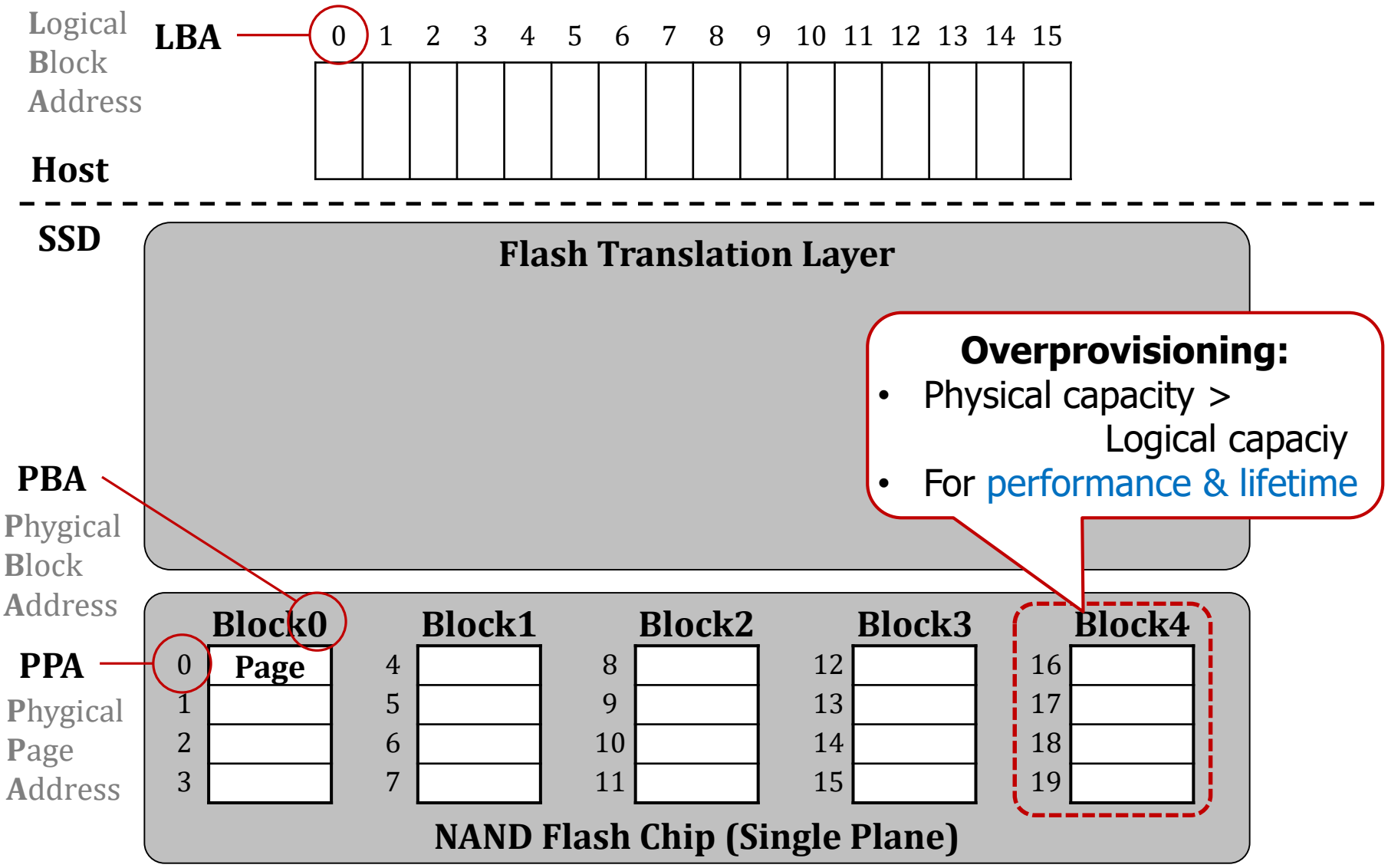
- SSD firmware (often referred to as SSD controller)
  - Provides backward compatibility with traditional HDDs
  - By hiding unique characteristics of NAND flash memory

- Responsible for many important SSD-management tasks
  - Address translation + garbage collection
    - Performs out-of-place writes due to erase-before-write property
  - Wear leveling
    - To prolong SSD lifetime by evenly distributing P/E cycles
  - Data refresh
    - Resets transient errors by copying data to a new page(s)
  - I/O scheduling
    - To take full advantage of SSD internal parallelism

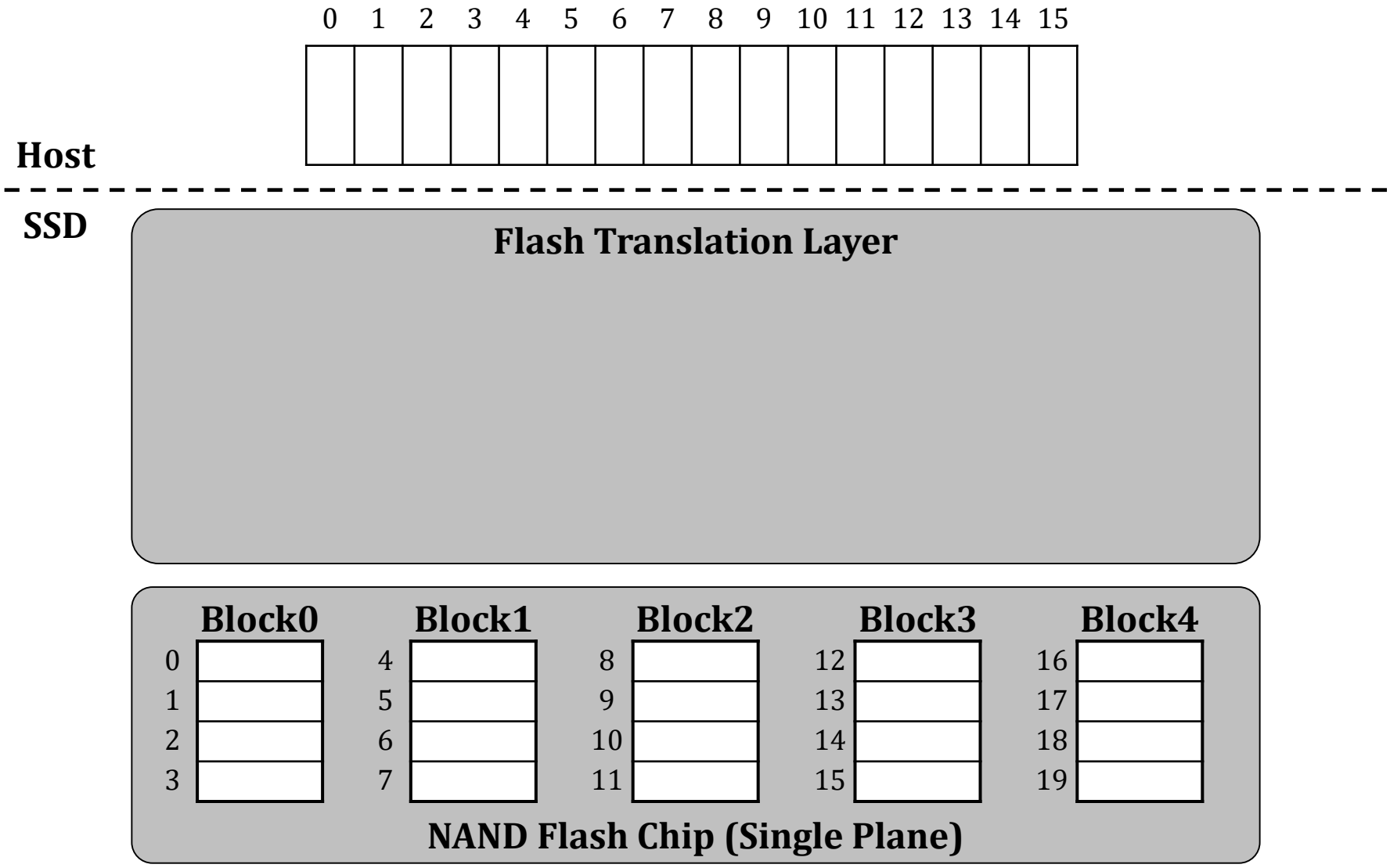
Simple SSD Architecture



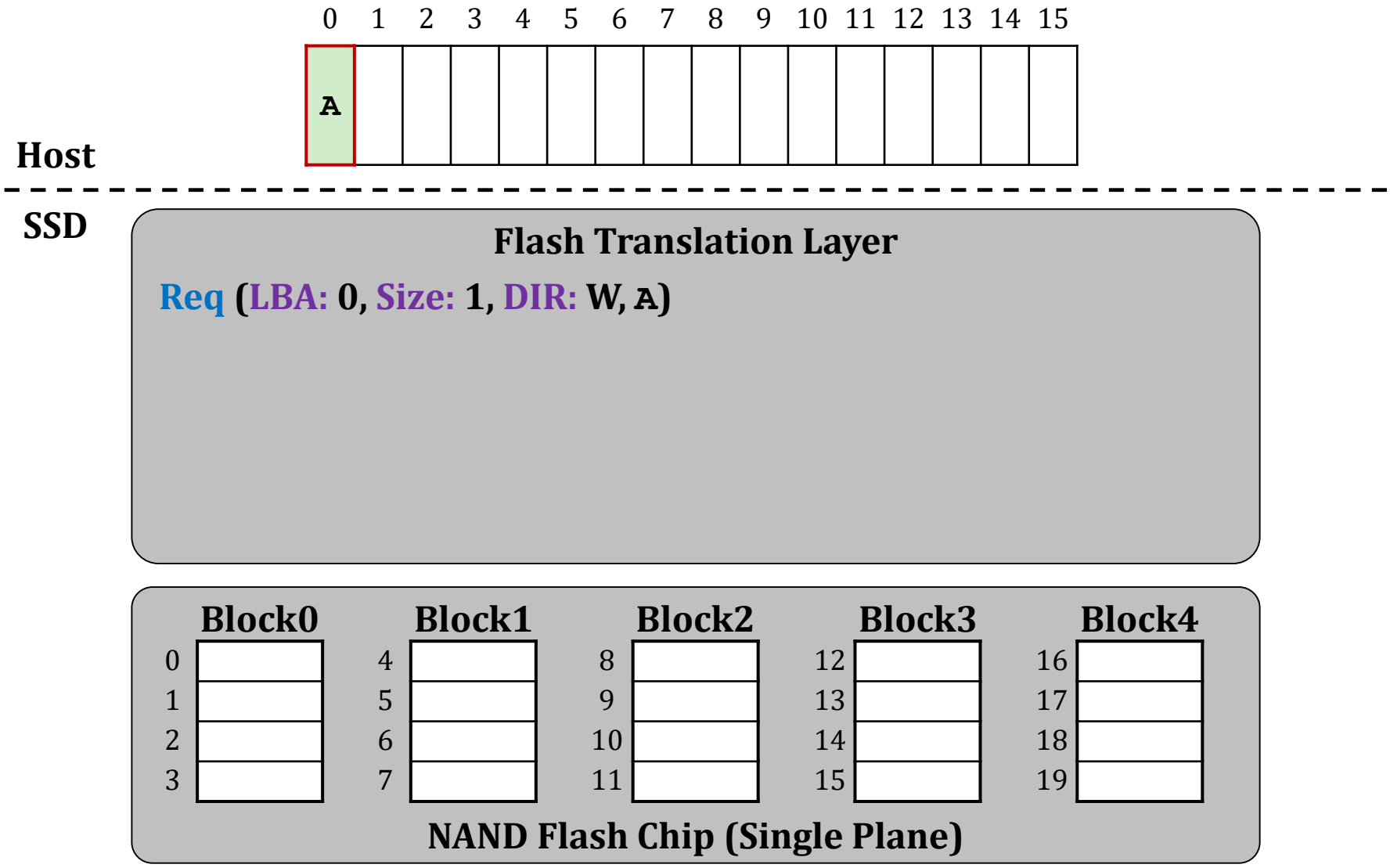
Simple SSD Architecture



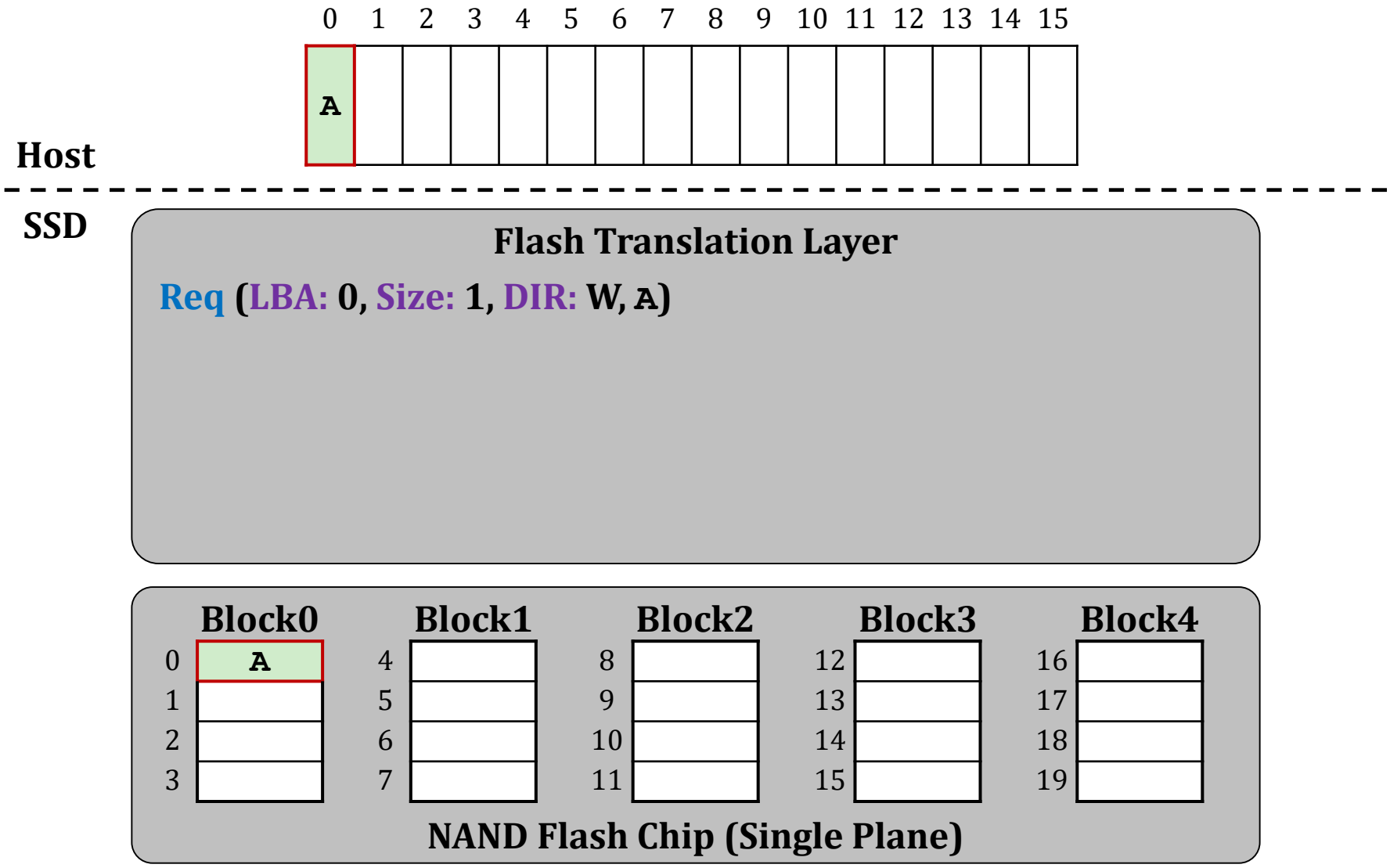
Write Request Handling: Page Write



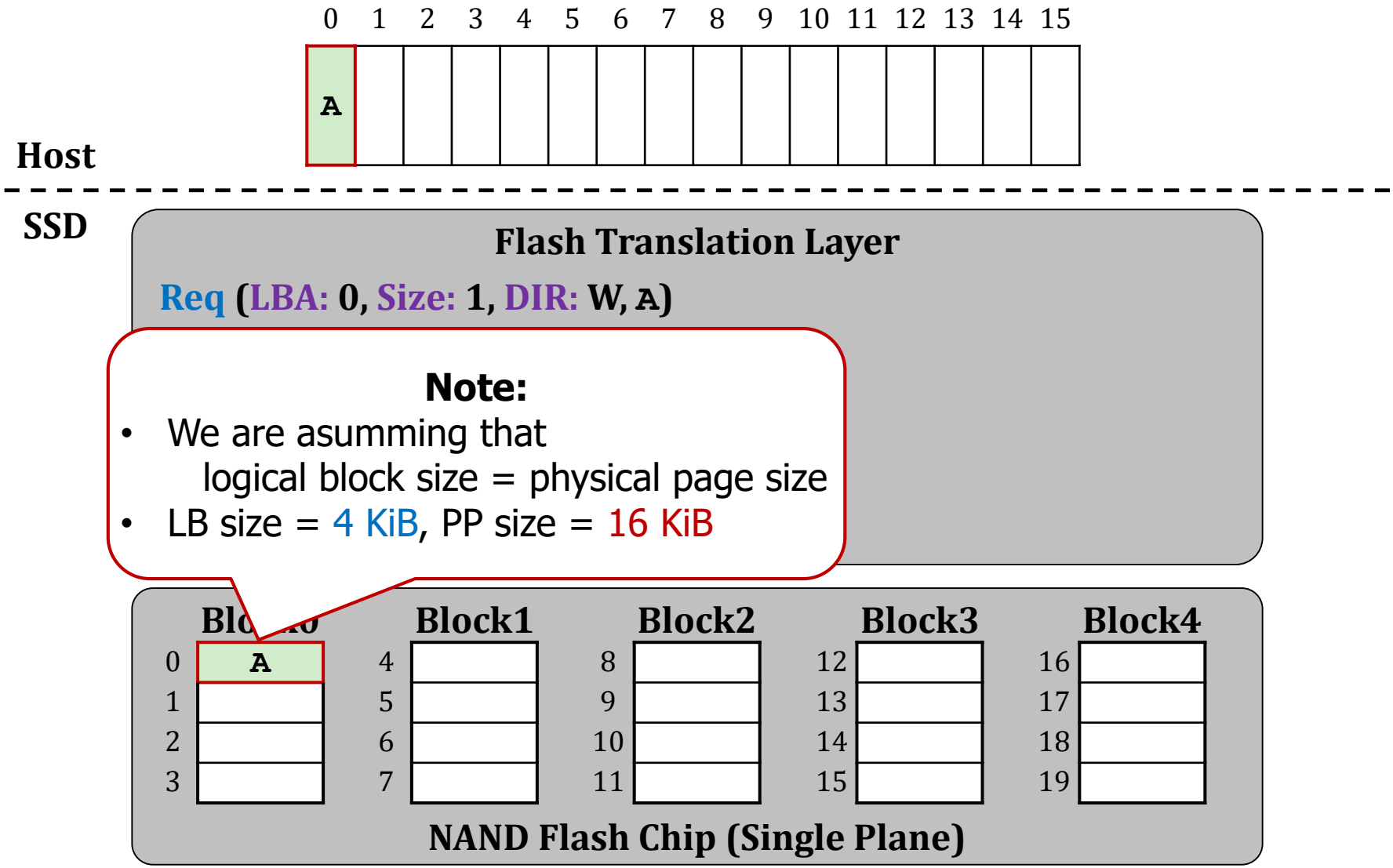
Write Request Handling: Page Write



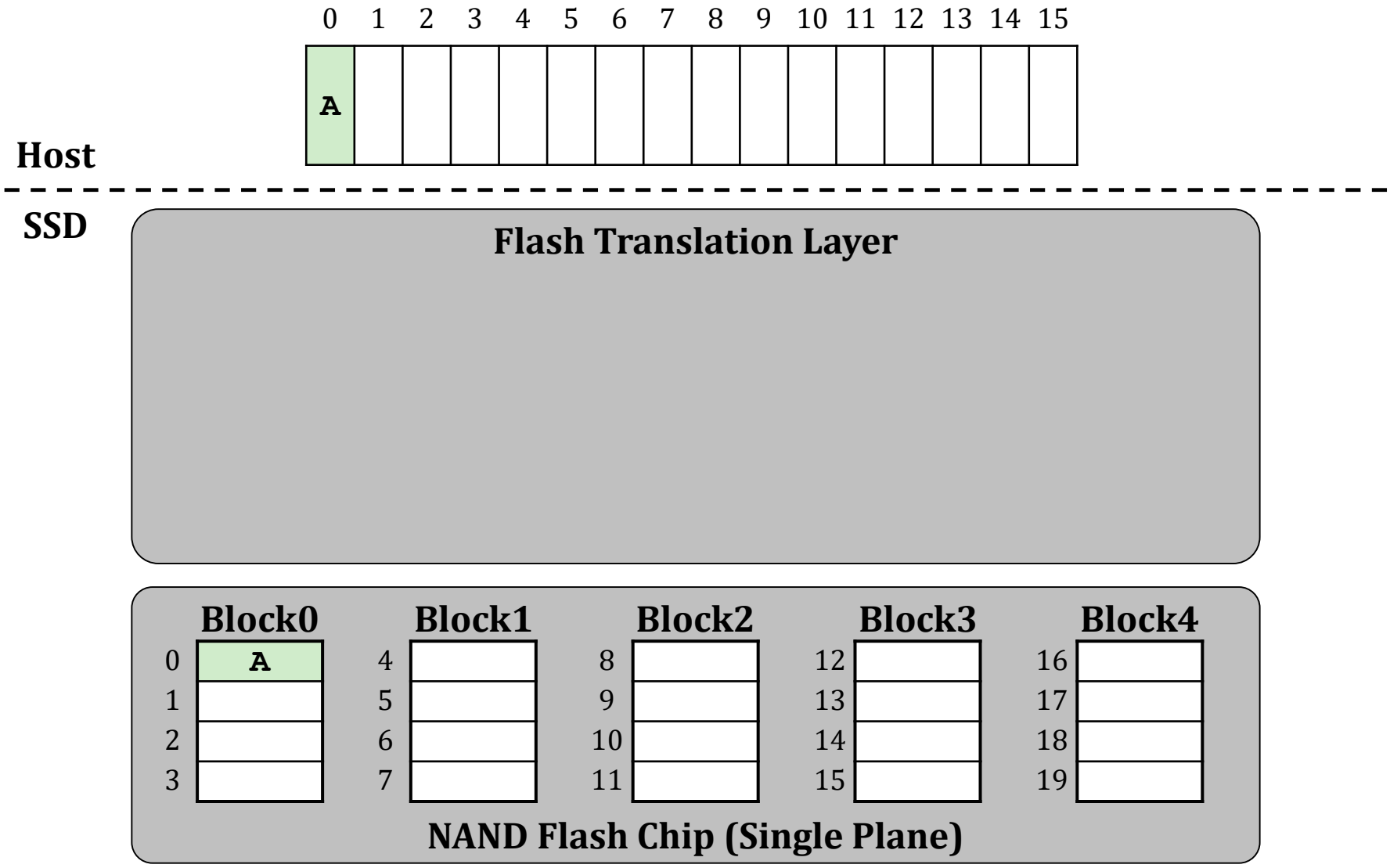
Write Request Handling: Page Write



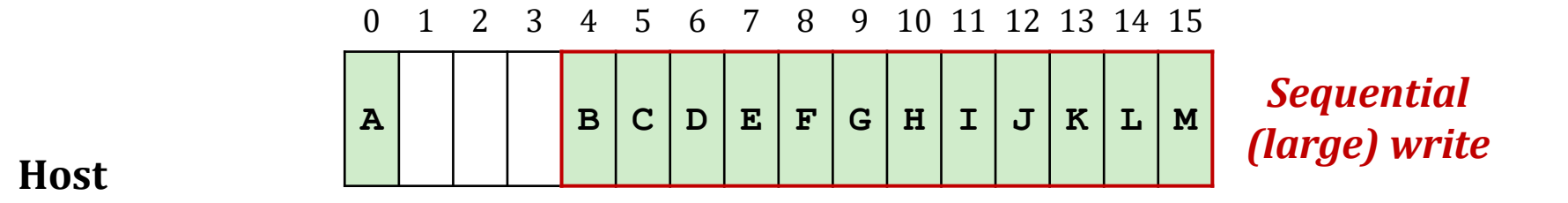
Write Request Handling: Page Write



Write Request Handling: Sequential Write



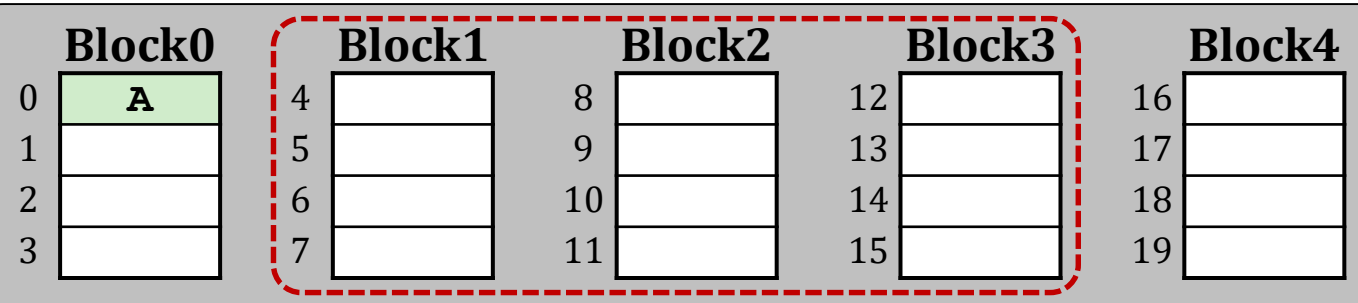
Write Request Handling: Sequential Write



SSD

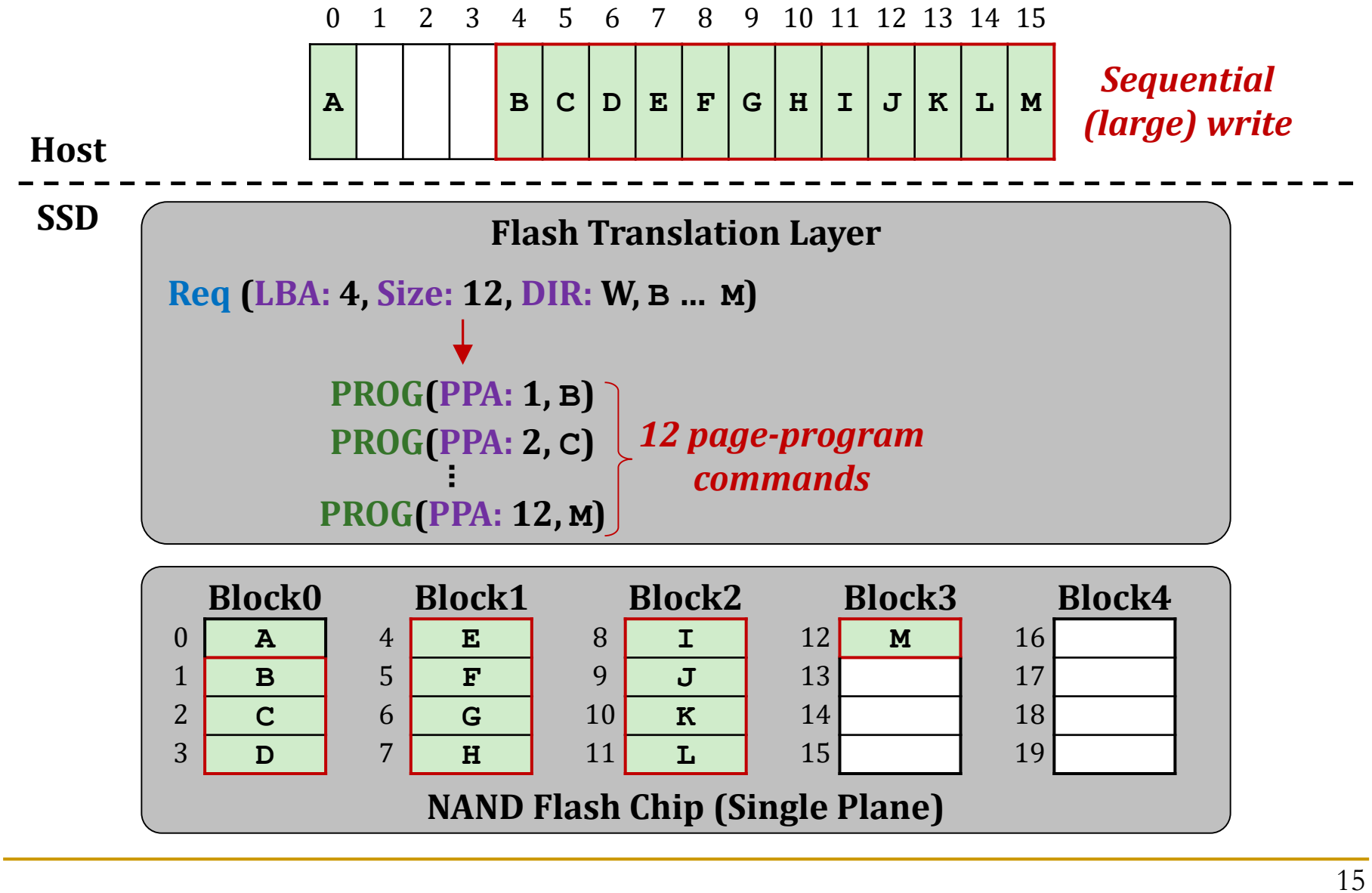
Flash Translation Layer

Req (LBA: 4, Size: 12, DIR: W, B … M)

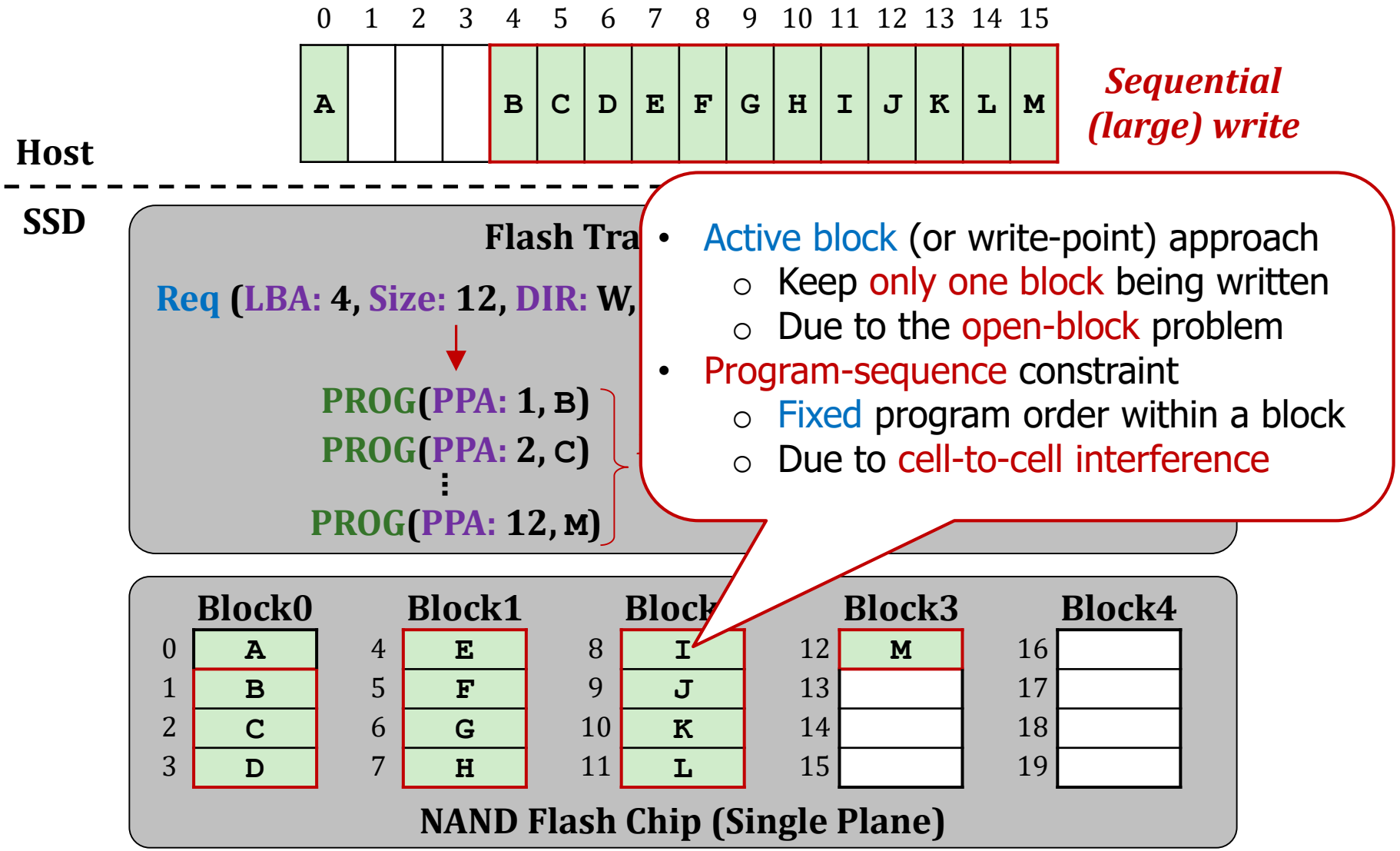


NAND Flash Chip (Single Plane)

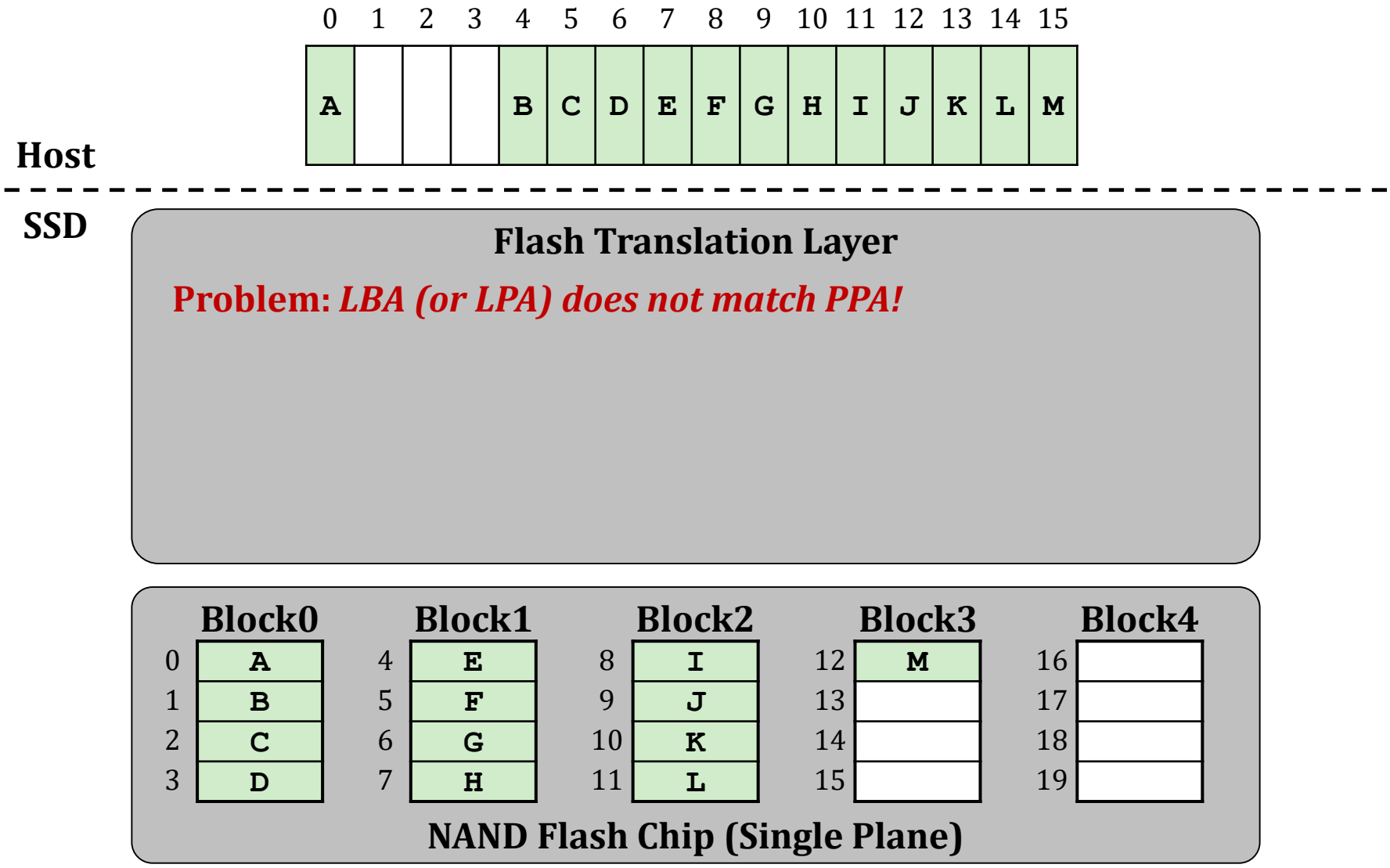
Write Request Handling: Sequential Write



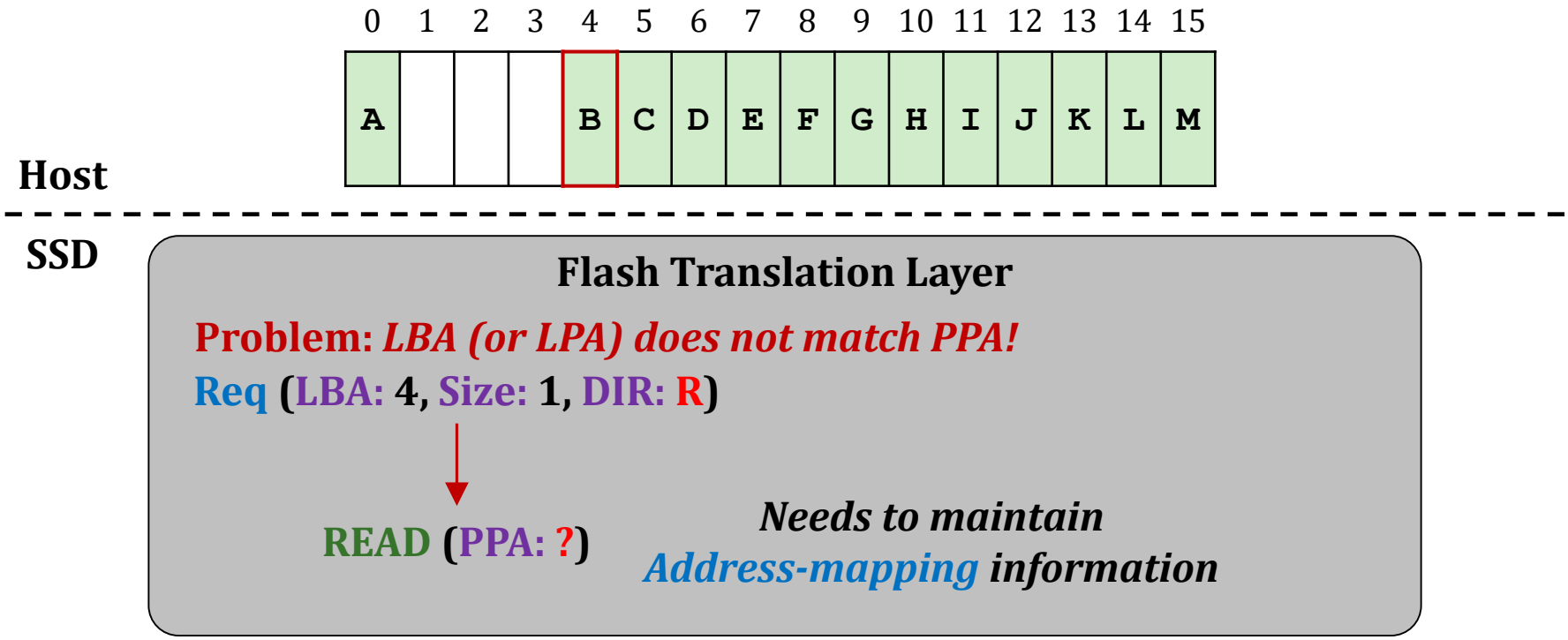
Write Request Handling: Sequential Write



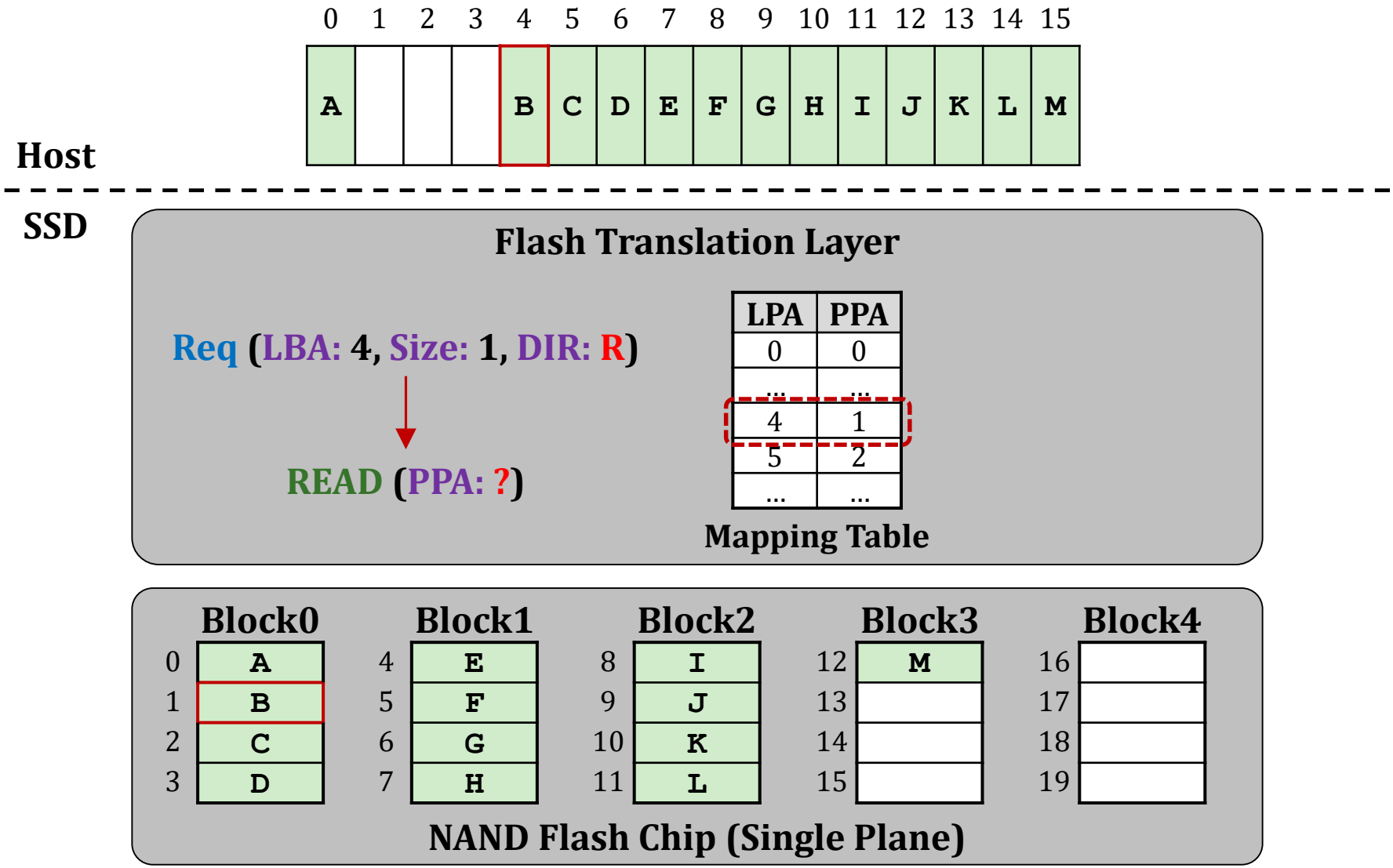
Write Request Handling: Address Mapping



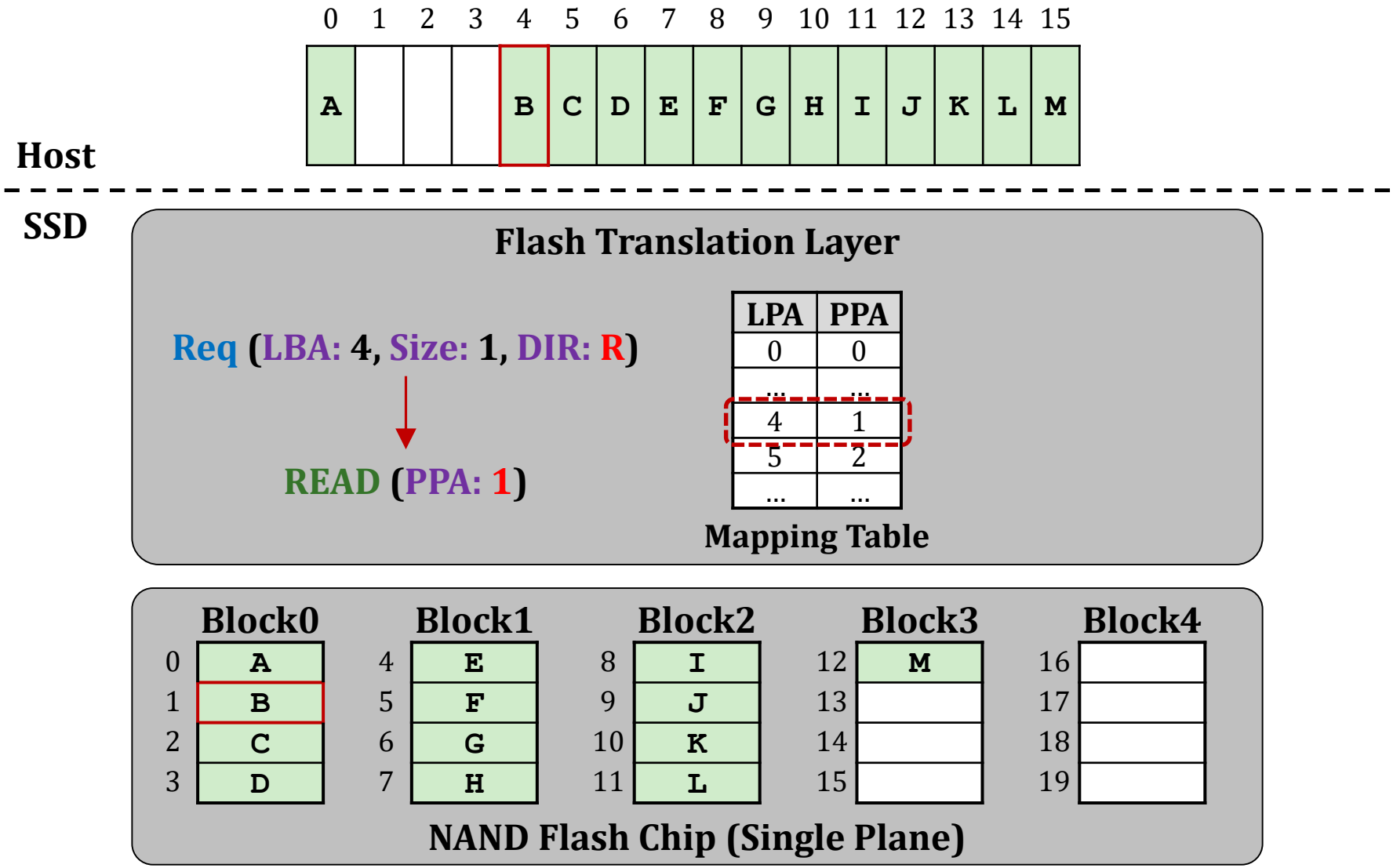
Write Request Handling: Address Mapping



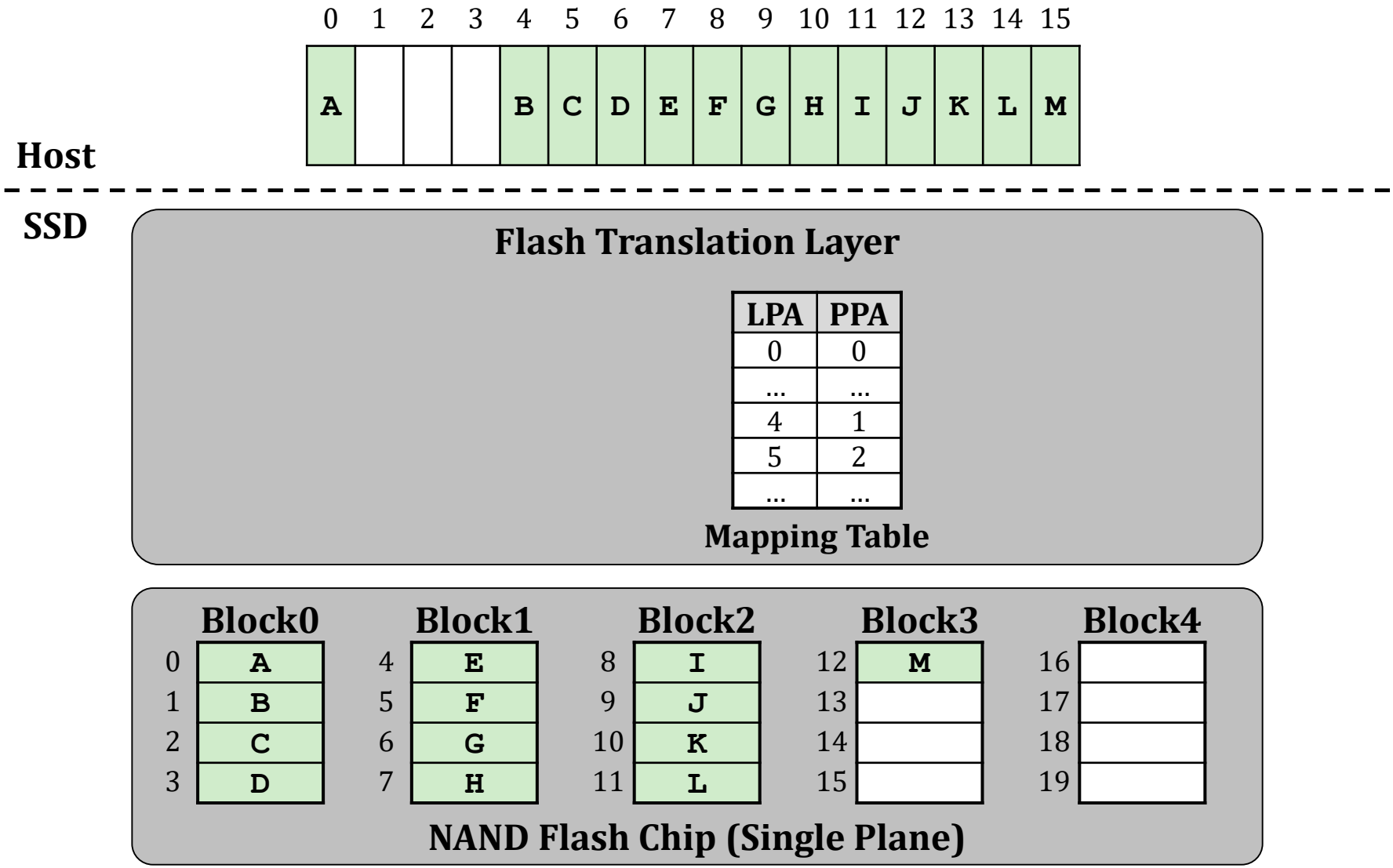
Write Request Handling: Address Mapping



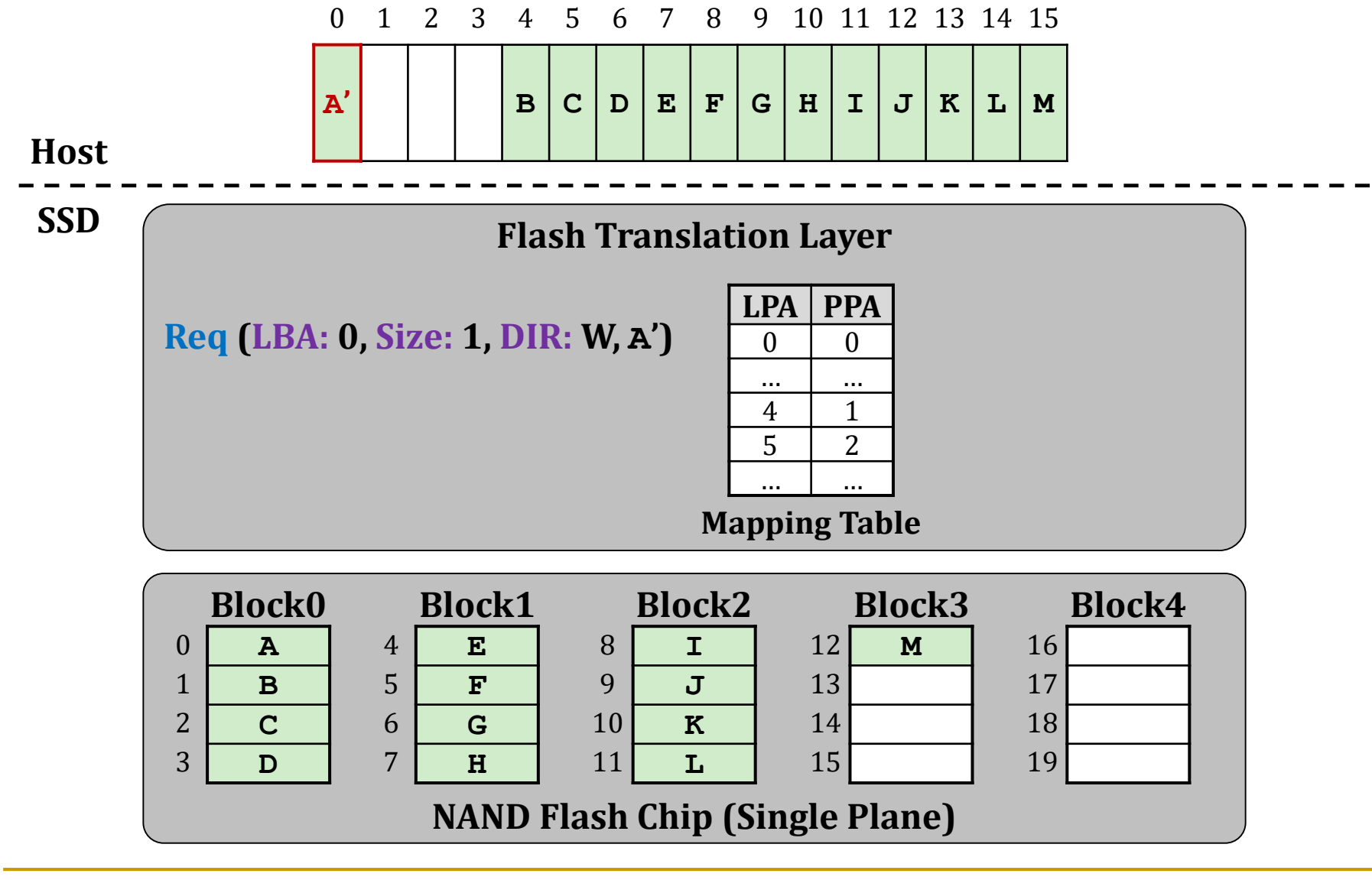
Write Request Handling: Address Mapping



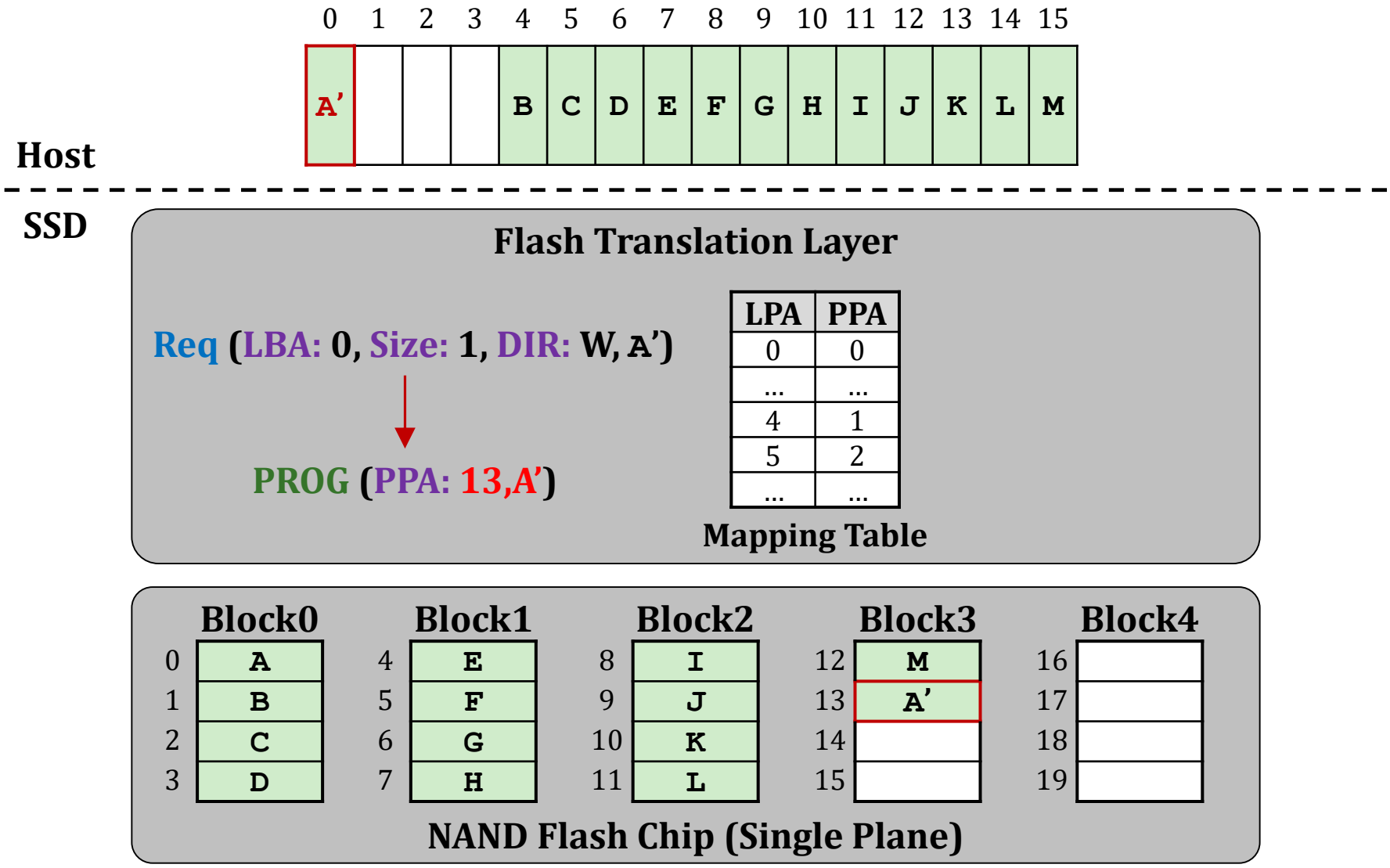
Write Request Handling: Update



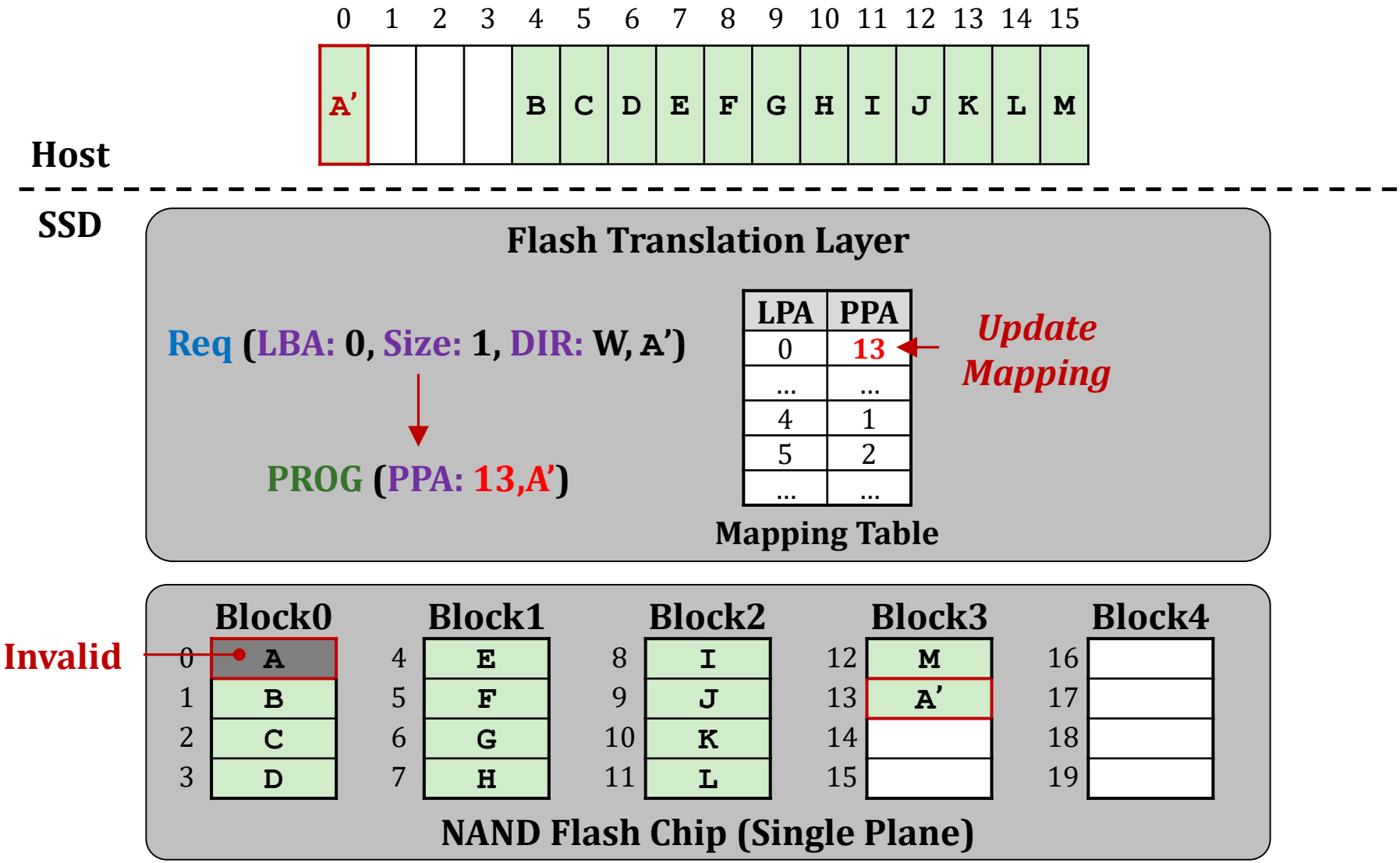
Write Request Handling: Update



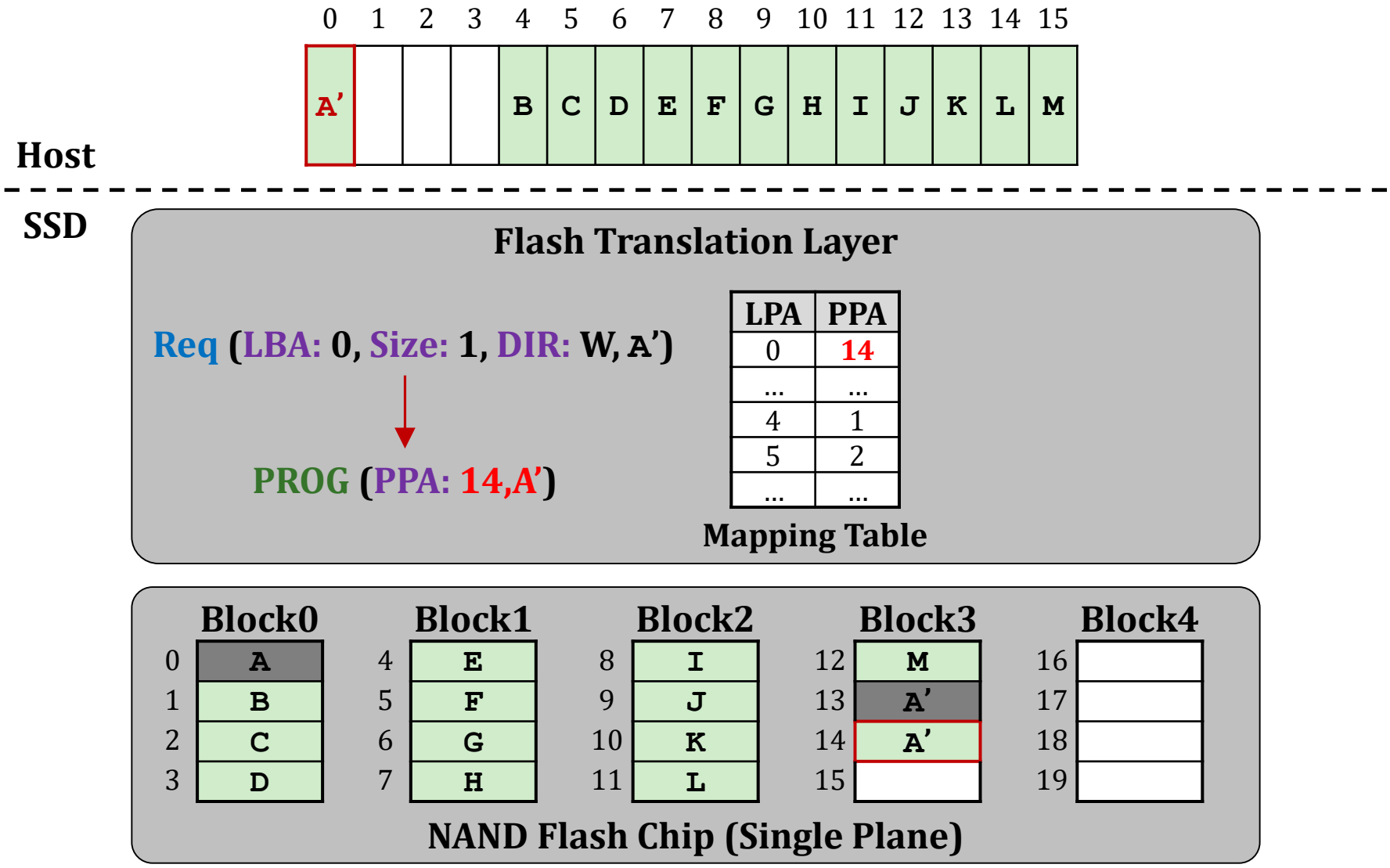
Write Request Handling: Update



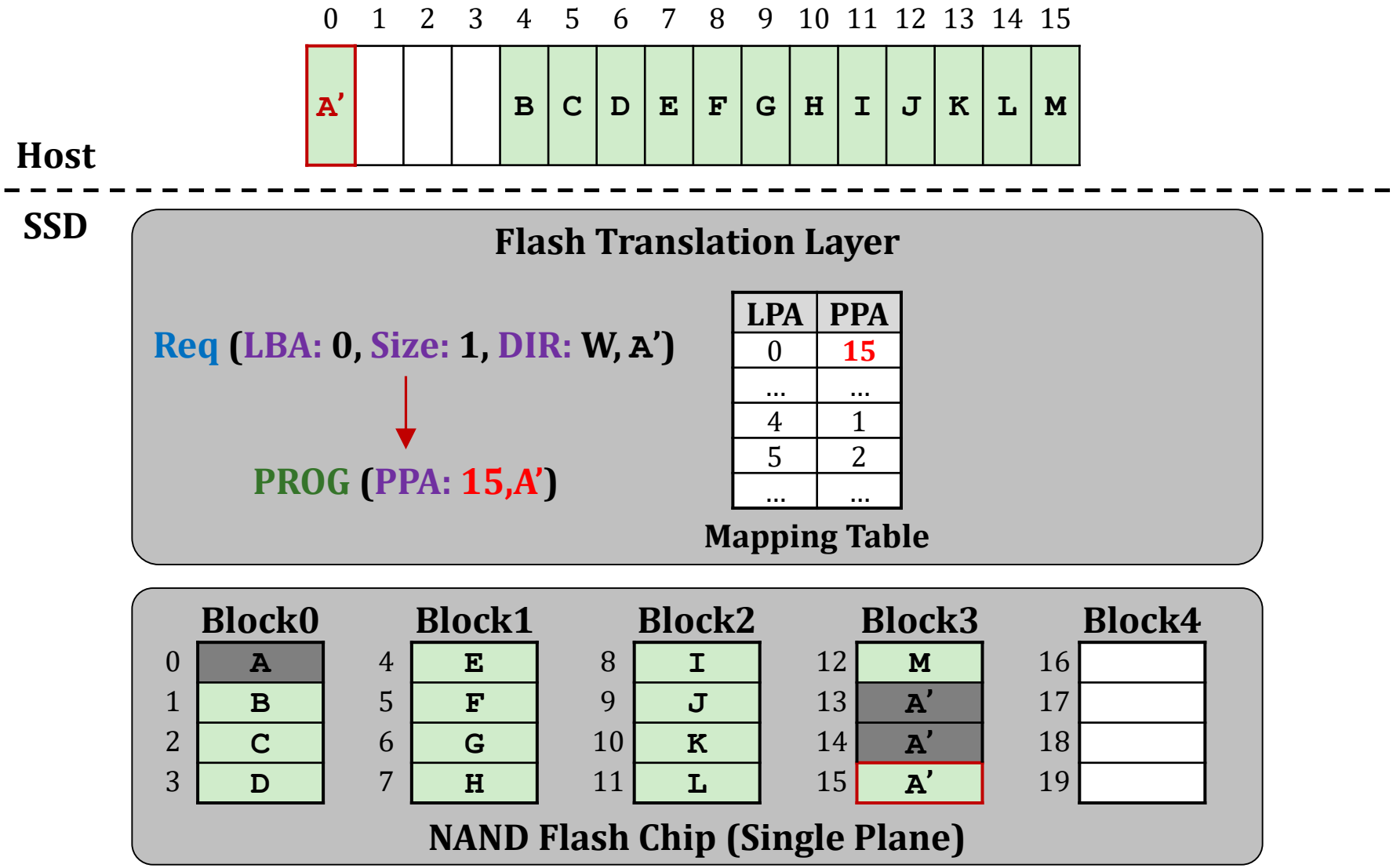
Write Request Handling: Update



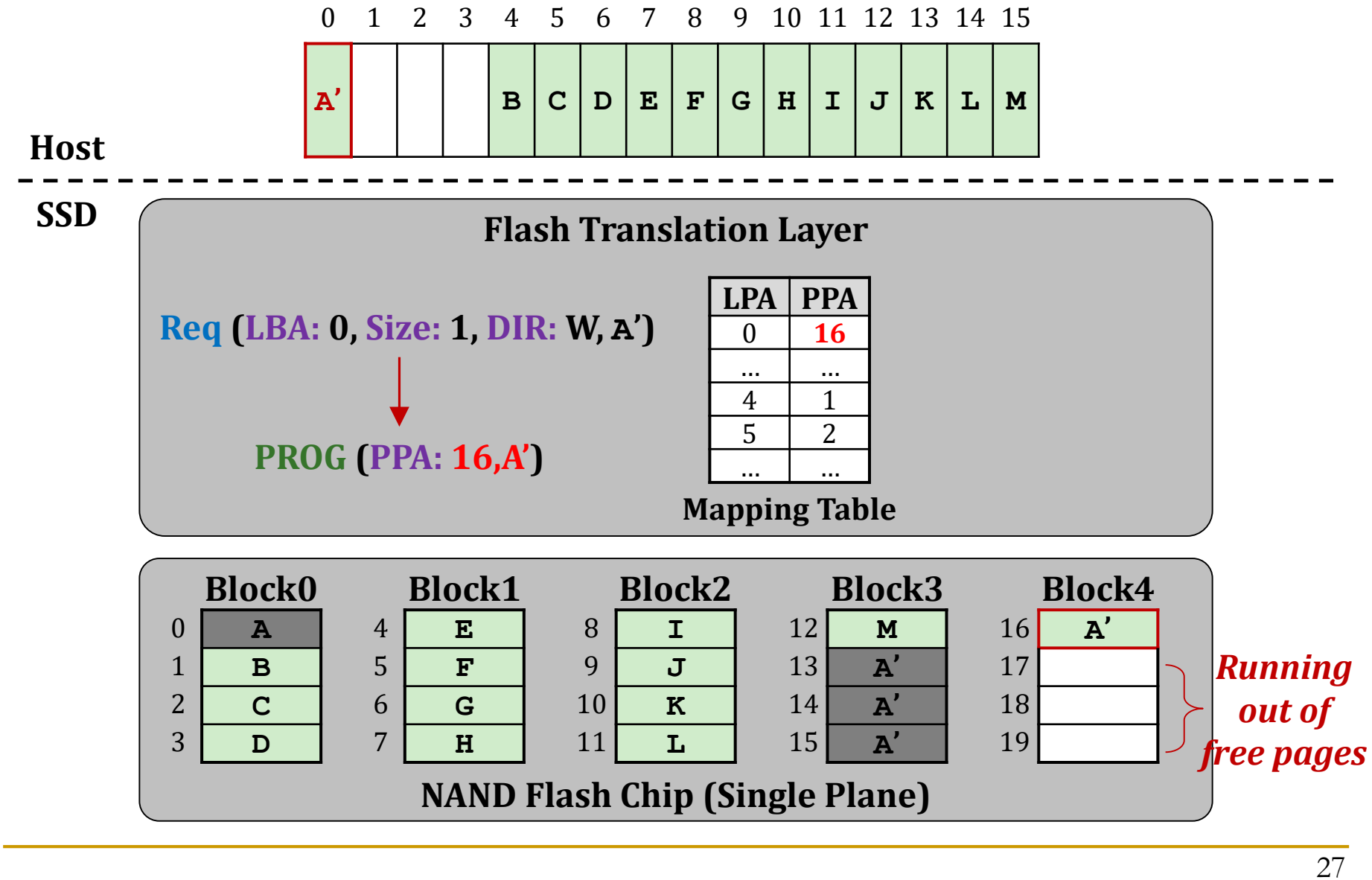
Write Request Handling: Update



Write Request Handling: Update



Write Request Handling: Update

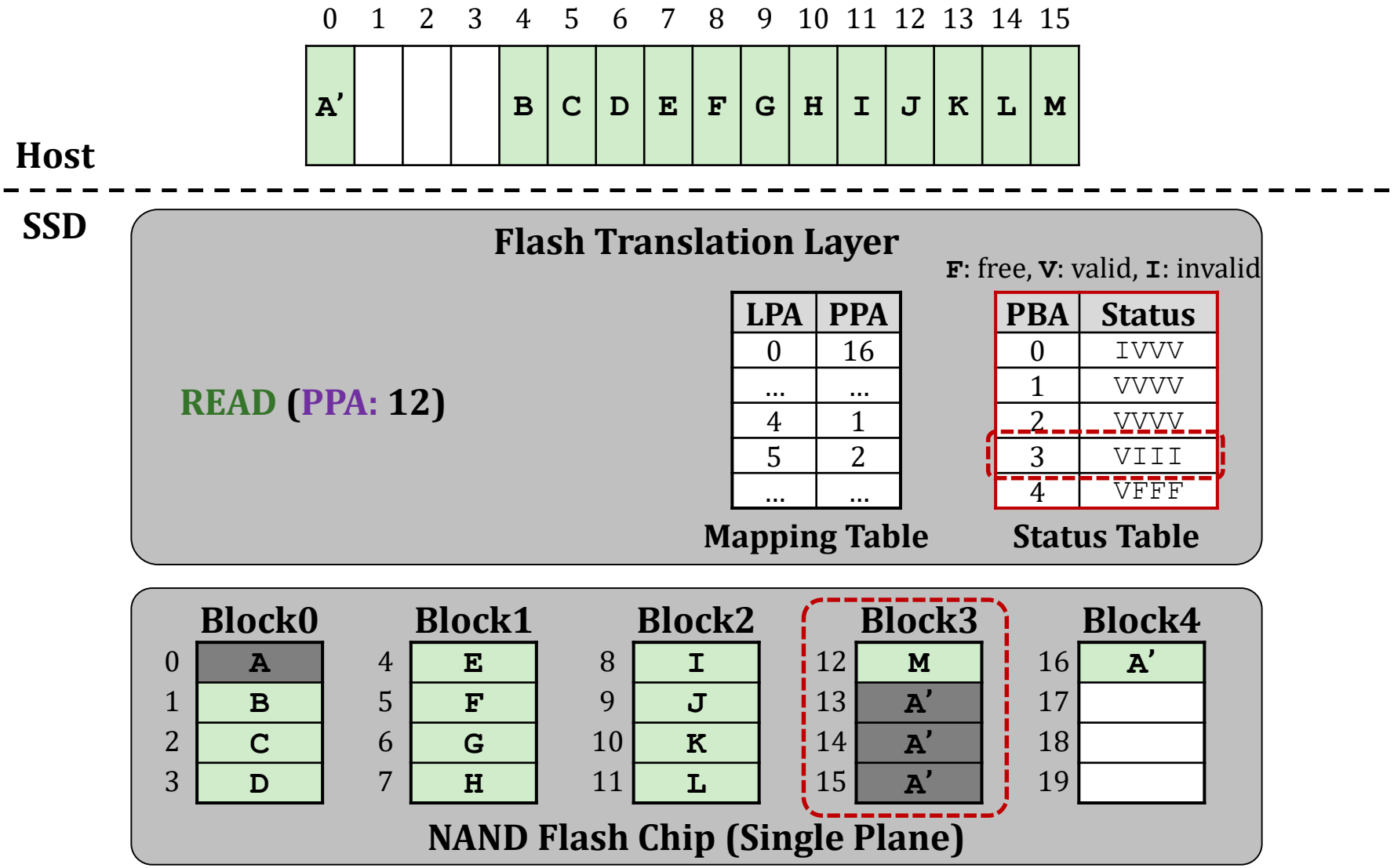


Garbage Collection

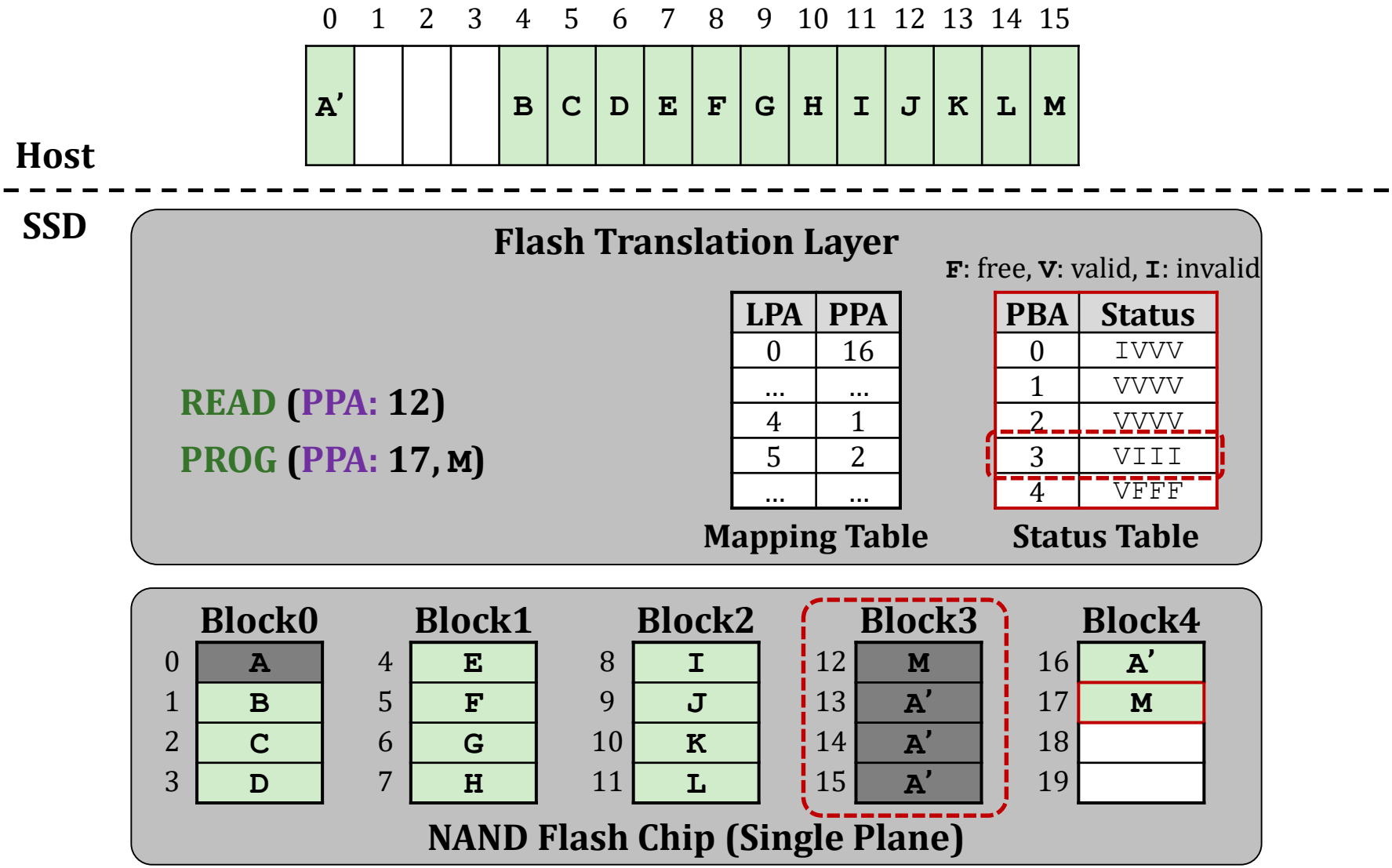
- Reclaims free pages by erasing invalid pages
  - Erase unit: block
  - If a victim block (to erase) has valid pages,
    all the valid pages need to be copied to other free pages
    - Performance overhead: (tREAD + tPROG) × # of valid pages
    - Lifetime overhead: additional writes → P/E-cycle increase

- Greedy victim-selection policy:
  Erases the block with the largest number of invalid pages
  - Needs to maintain # of invalid (or valid) pages for each block

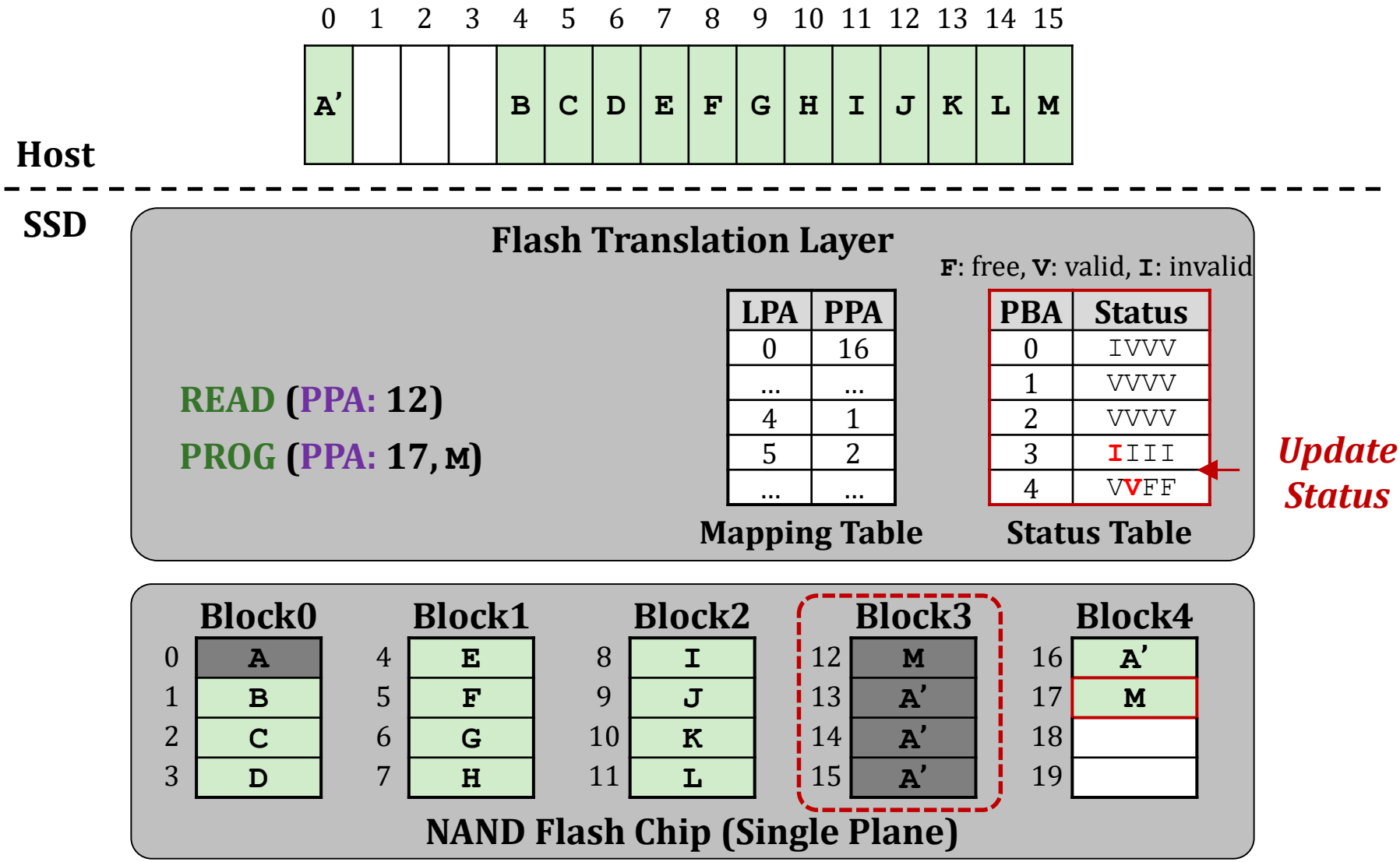
Write Request Handling: Garbage Collection



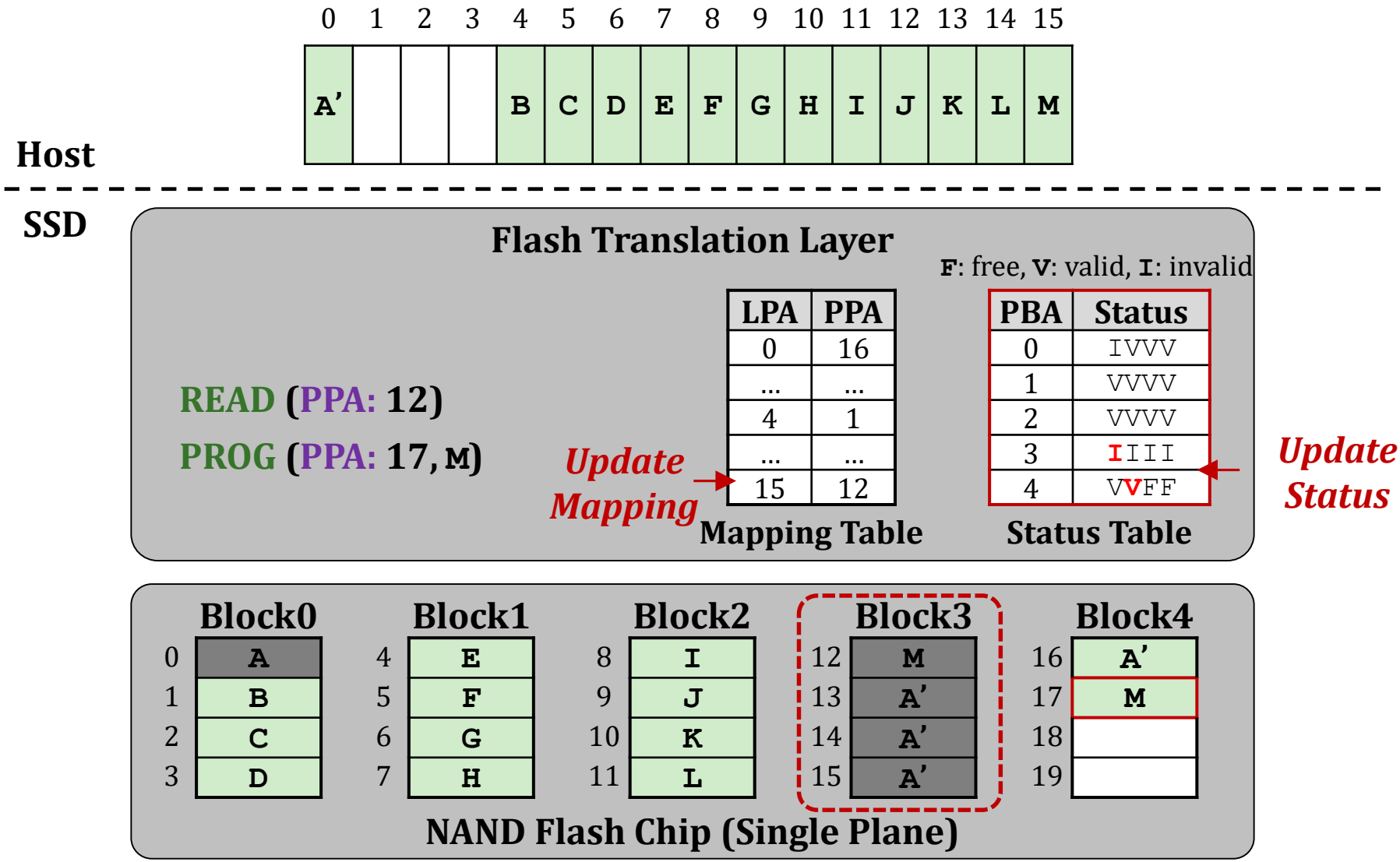
Write Request Handling: Garbage Collection



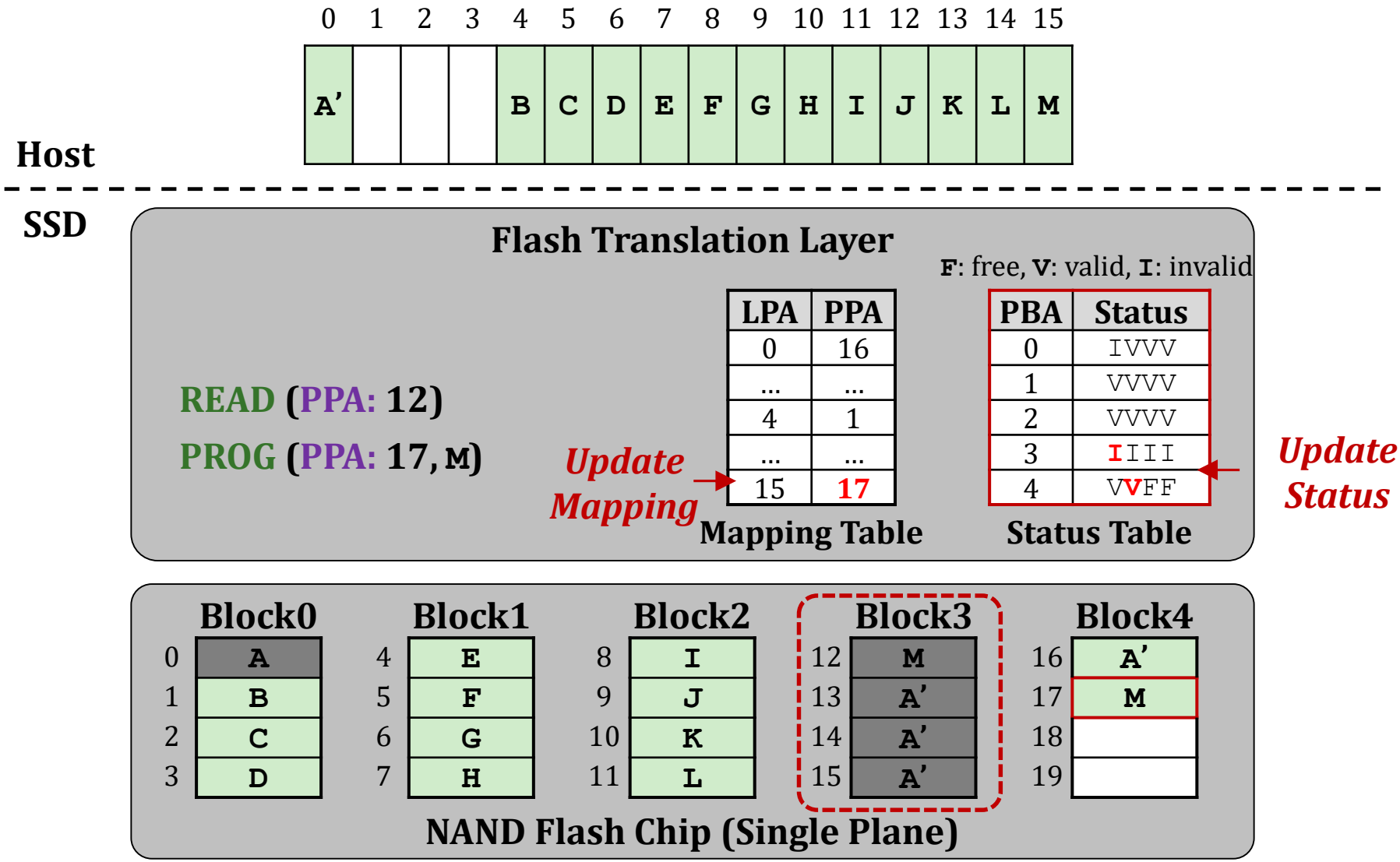
Write Request Handling: Garbage Collection



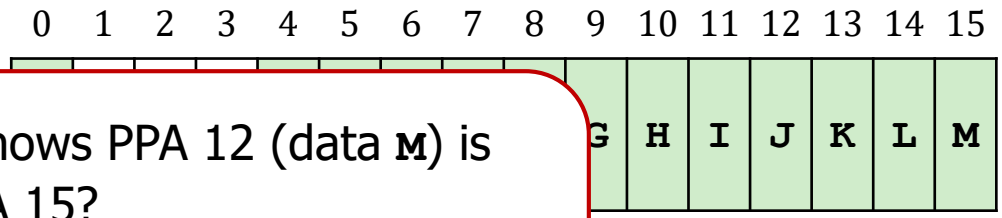
Write Request Handling: Garbage Collection



Write Request Handling: Garbage Collection



Write Request Handling: Garbage Collection



- **Q:** How FTL knows PPA 12 (data **M**) is mapped to LPA 15?
  - Unless it maintains P2L mappings?
- **A:** P2L mapping is stored in each physical page's OOB (Out-of-Band) area

READ (PPA: 12)

PROG (PPA: 17, M)

on Layer

F: free, **V**: valid, **I**: invalid

LPA	PPA
0	16
...	...
4	1
...	...
15	17

Mapping Table

Update Mapping

PBA	Status
0	IVVV
1	VVVV
2	VVVV
3	IIII
4	VVFF

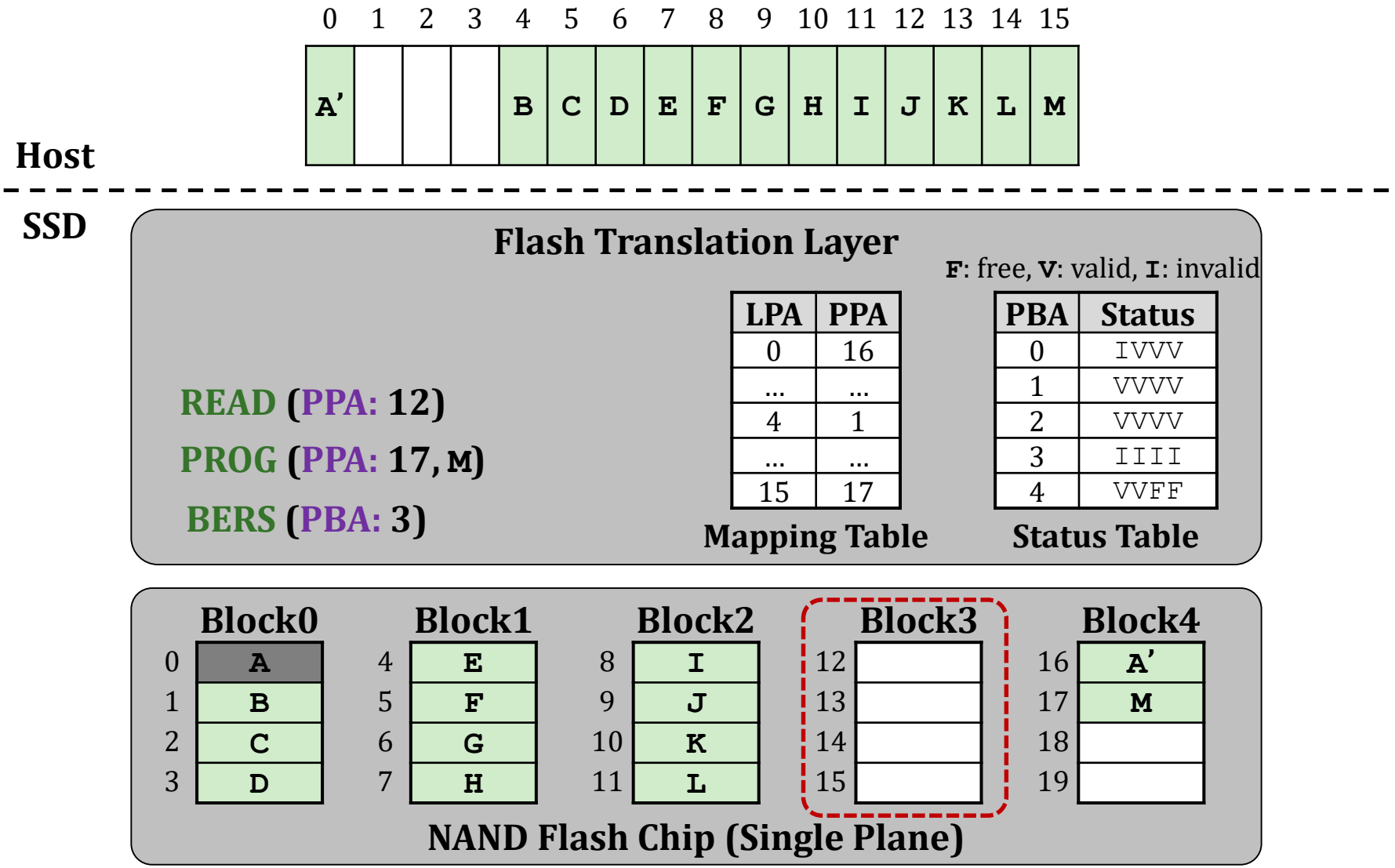
Status Table

Update Status

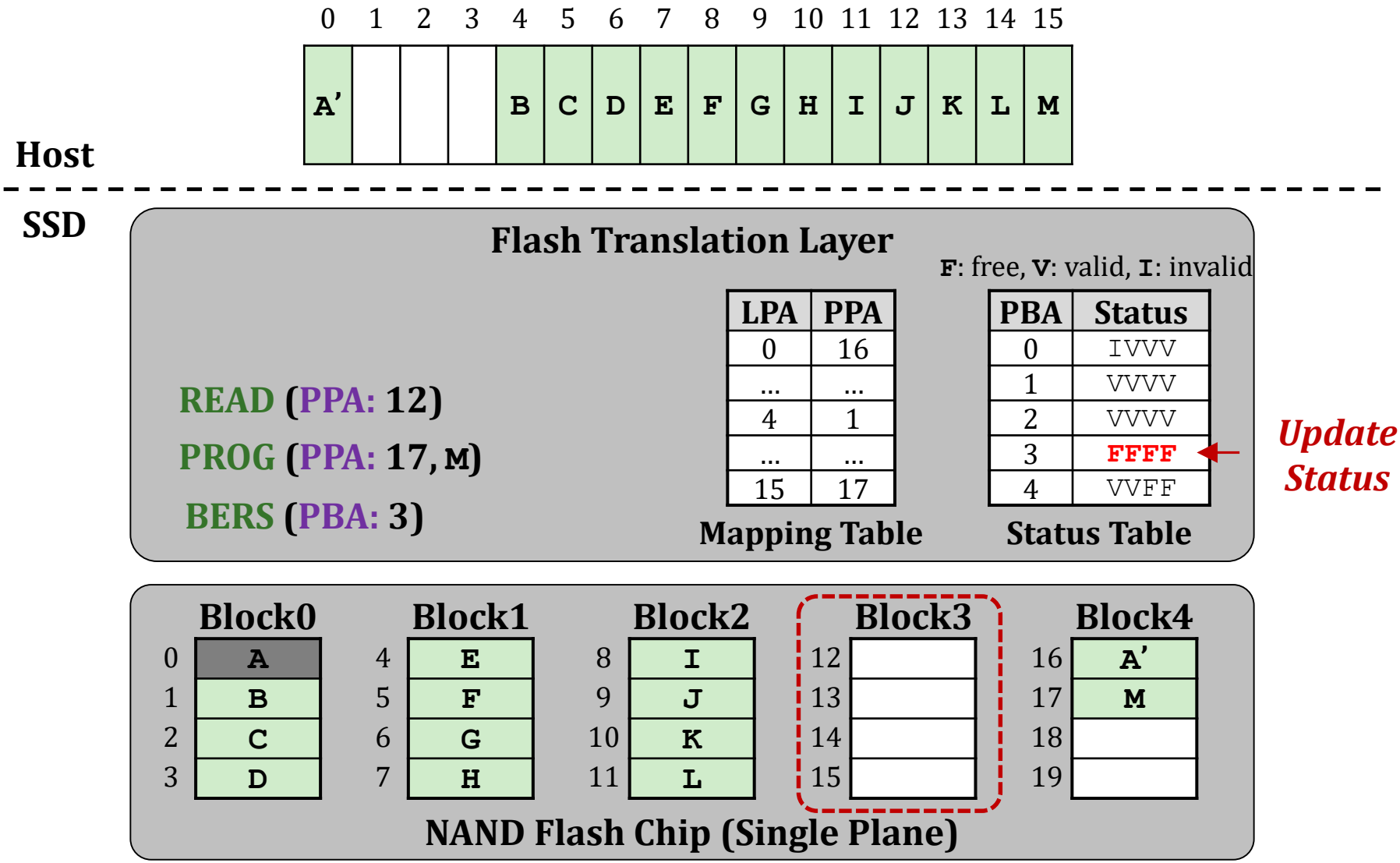
Block0		Block1		Block2		Block3		Block4	
0	A	4	E	8	I	12	M	16	A'
1	B	5	F	9	J	13	A'	17	M
2	C	6	G	10	K	14	A'	18	
3	D	7	H	11	L	15	A'	19	

NAND Flash Chip (Single Plane)

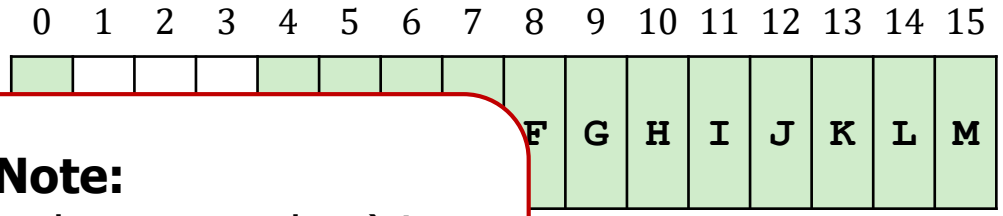
Write Request Handling: Garbage Collection



Write Request Handling: Garbage Collection



Write Request Handling: Garbage Collection



Note:

- Block erasure (and status update) is done just before programming a new page to the block (i.e., lazy erase)
  - Due to the open-block problem

ation Layer

F: free, **V**: valid, **I**: invalid

LPA	PPA
0	16
...	...
4	1
...	...
15	17

Mapping Table

PBA	Status
0	IVVV
1	VVVV
2	VVVV
3	FFFF
4	VVFF

Status Table

Update Status

(PPA: 12)

G (PPA: 17, M)

BERS (PBA: 3)

Block0		Block1		Block2		Block3		Block4	
0	A	4	E	8	I	12		16	A'
1	B	5	F	9	J	13		17	M
2	C	6	G	10	K	14		18	
3	D	7	H	11	L	15		19	

NAND Flash Chip (Single Plane)

Performance Issues

- Garbage collection significantly affects SSD performance
  - High latency: Large block size of modern NAND flash memory
    - Assume 1) a block contains 576 pages,
      2) only 5% of the pages in the victim block are valid
      3) tR = 100 us, tPROG = 700 us, tBERS = 5 ms
      - # of pages to copy = $576 \times 0.05 = 28.8 \rightarrow 28$ pages
      - GC latency > $28 \times (tR + tPROG) + tBERS = 27{,}400$ us
    - Order(s) of magnitude larger latency than tR and tPROG
    - Copy operations are the major contributor (rather than tBERS)
  - If FTL performs GC in an atomic manner,
    it delays user requests for a significantly long time
    - Long tail latency (performance fluctuation)
    - Noisy neighbor: a read-dominant workload's performance would be significantly affected when running with a write-intensive workload (+ performance fairness problem)

Performance Issues: Mitigation

- TRIM (UNMAP or discard) command
  - Informs FTL of deletion/deallocation of a logical block
  - Allows FTL to skip copy of obsolete (i.e., invalid) data

- Background GC: Exploits SSD idle time
  - Challenge: how to accurately predict SSD idle time
  - Premature GC: copied pages could have been invalidated by the host system

- Progressive GC: Divide GC process into subtasks
  - e.g., copying 28 pages → (copying 1 page + servicing user request) × 28
  - Effective at decreasing tail latency

Required Materials

- Address Mapping
  - Aayush Gupta, Yongjae Kim, and Bhuvan Urgaonkar, "DFTL: A Flash Translation Layer Employing Demand-based Selective Caching of Page-level Address Mappings," In ASPLOS 2009.

Recommend Materials

- Cache read & Read-retry
  - Jisung Park, Myungsuk Kim, Lois Orosa, Jihong Kim, and Onur Mutlu, "Reducing Solid-State Drive Read Latency by Optimizing Read-Retry," In ASPLOS 2021.

- Program & Erase Suspension
  - Guanying Wu and Xunbin He, "Reducing SSD Read Latency via NAND Flash Program and Erase Suspension," In USENIX FAST 2012.
  - Shine Kim, Jonghyun Bae, Hakbeom Jand, Wenjing Jin, Jeonghun Gong, Seungyeon Lee, Tae Jun Ham, and Jae W. Lee, "Practical Erase Suspension for Modern Low-latency SSDs," In USENIX ATC 2019.

P&S Modern SSDs

Address Mapping & Garbage Collection

Dr. Mohammad Sadrosadati
Prof. Onur Mutlu
ETH Zürich
Fall 2022
23 November 2022